LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-TR-678127

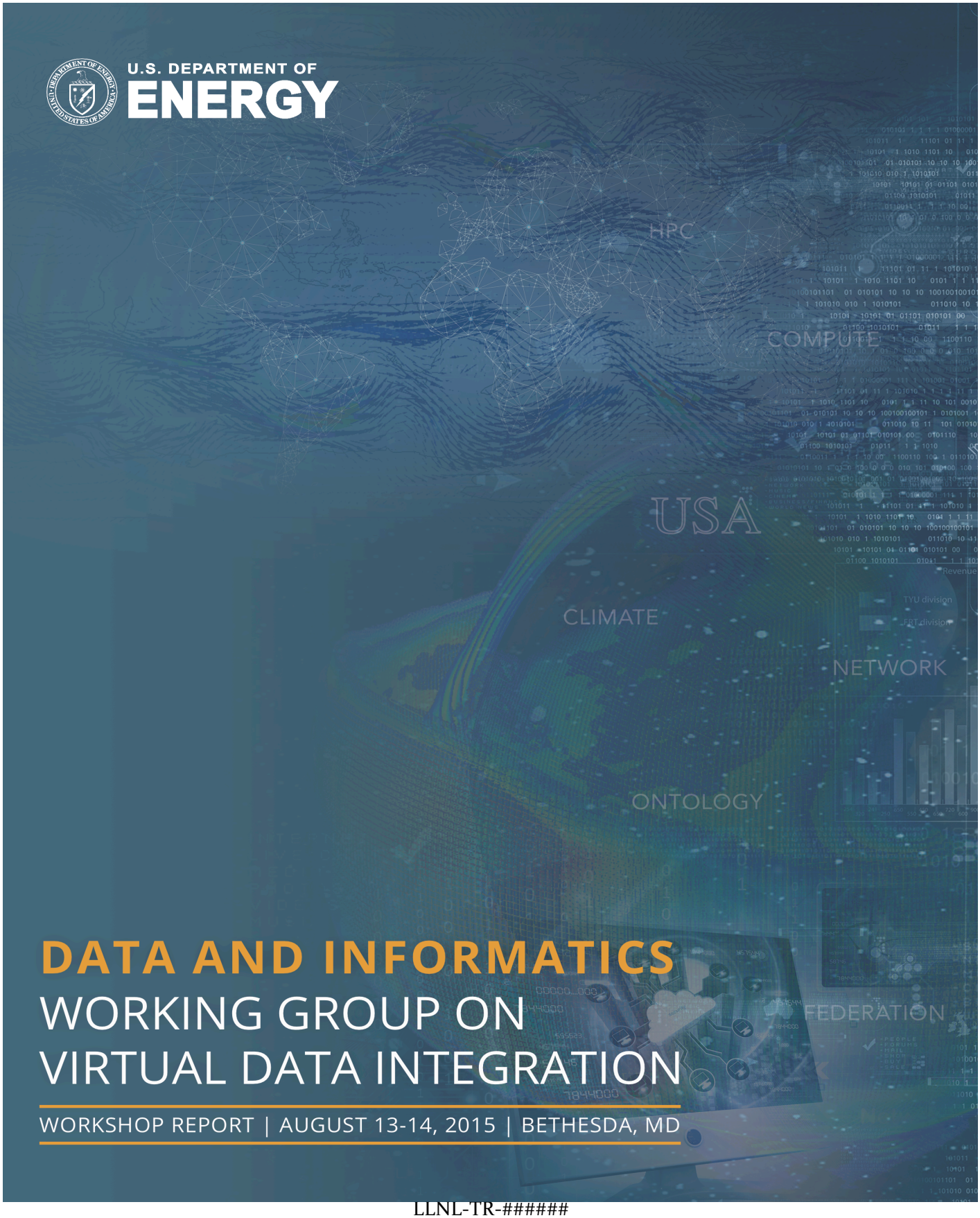
DATA AND INFORMATICS WORKING GROUP ON VIRTUAL DATA INTEGRATION WORKSHOP REPORT

D. N. Williams, G. Palanisamy, K. Kleese-Van Dam

October 13, 2015

Disclaimer

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.



DATA AND INFORMATICS
WORKING GROUP ON VIRTUAL DATA INTEGRATION

WORKSHOP REPORT | AUGUST 13-14, 2015 | BETHESDA, MD

LLNL-TR-######

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

DATA AND INFORMATICS WORKING GROUP ON VIRTUAL DATA INTEGRATION WORKSHOP REPORT

AUGUST 13-14, 2015

BETHESDA, MD

Requirements to Achieve BER's Vision of a Virtual Laboratory

Next Generation Data Infrastructure for Climate Science

Office of Biological and Environmental Research

Program Manager: Dr. Justin Hnilo

Workshop Attendees and Report Contributors:

Deb Agarwal, David Bader, Tom Boden, Scott Collis, Jennifer Comstock, Eli Dart, Paul Durack, Ian Foster, Forrest Hoffman, Robert Jacob, Phil Rasch, Timothy Scheibe, Mallikarjun Shankar, David Skinner, Peter Thornton, Margaret Torn, Andrew Vogelmann, Michael Wehner, Shaocheng Xie

Workshop and Report Organizers:

Kerstin Kleese-Van Dam	PNNL/BNL	Kleese@bnl.gov	631.344-6019
Giriprakash Palanisamy	ORNL	palanisamyg@ornl.gov	865.241-5926
Dean N. Williams	LLNL	Williams13@llnl.gov	925.423-0145

Table of Contents

Workshop Narrative

1. Executive Summary

This report is the outcome of a workshop that was commissioned by the Department of Energy's Climate and Environmental Sciences Division (CESD) to examine current and future data infrastructure requirements that would be foundational to achieving CESD's scientific mission goals. Over the past several years, data volumes in CESD disciplines have risen sharply to unprecedented levels (tens of petabytes). So too has the complexity and diversity of the research data (simulation, observation, and reanalysis) needing to be captured, stored, verified, analyzed, and integrated. With the trends of increased data volume (in the hundreds of petabytes), more complex analysis processes, and growing cross-disciplinary collaborations, it is timely to investigate whether the CESD community has the right computational and data support to realize the full scientific potential from its data collections. In recognition of the challenges, a partnership is forming across CESD and with national and international agencies to investigate the viability of creating an integrated, collaborative data infrastructure: a *virtual laboratory*. The overarching goal of this report is to identify the community's key data technology *requirements* and high-priority *development needs* for sustaining and growing their scientific discovery potential. The report also aims to map these requirements to *existing solutions* and to identify *gaps* in current services, tools, and infrastructure that will need to be addressed in the short, medium, and long term so as not to impede scientific progress.

Prior to the workshop, a survey was circulated to attendees and their associates. Responses emphasized, in particular, a concern about sustained supply of sufficient computational and storage resources. More broadly, they indicated a need for cross-cutting integrating solutions that address the full spectrum of data lifecycle issues—collection, management, annotation, analysis, sharing, visualization, workflows, and provenance. The following were the top-ten most cited requirements: (1) an easy way to publish and archive data; (2) comparison of heterogeneous data types; (3) user support and documentation; (4) access to observational and experimental resources; (5) scientific and computational reproducibility; (6) data movement from archive to supercomputers; (7) unifying single user accounts across DOE resources and facilities; (8) reliability and resiliency of resources; (9) intuitive human-computer interaction; and (10) quality control algorithms for data. In addition, methodologies for knowledge gathering, management, and sharing were seen as an overall area that requires more community attention.

In addition to the survey, the report recognizes community infrastructure investments that support and enable analysis of massive, distributed scientific data collections and that leverage distributed architectures and compute environments designed for specific needs. The report captures this trend by first recognizing the scientific challenges in the form of diverse and disparate use cases. These scientific use cases capture and emphasize the need for data services, data centers, interoperable services, advanced computational environments, data analytics, data monitoring, multi-agency collaboration, and the evaluation of existing tools and services for potential reuse. Workshop discussion of community infrastructures to help build CESD's Virtual Laboratory include the Earth System Grid Federation (ESGF), which primarily serves simulation data to the global climate research community; the Atmospheric Radiation Measurement (ARM) Climate Research Facility's data center, which collects and serves observational instrument data; and the Carbon Dioxide Information and Analysis Center (CDIAC), which contains observations of ecosystem level exchanges of CO_2, water, energy, and momentum at different time scales for sites in the Americas. However, while these infrastructures may cover some of the requirements of the scientific use cases, they are lacking in generality in addressing all the use cases and will require enhancements to fulfill CESD's scientific vision.

The workshop itself produced the following core findings:

- The wider CESD community identified knowledge capture, management, and sharing as a key development area.
- The CESD community identified requirements for additional enabling data capabilities throughout the full research lifecycle from data discovery, multi-source data treatment handling large data volumes, flexible data analysis tools, reproducibility, and data publication and attribution.
- Attendees of the workshop voiced strong concerns regarding the lack of storage and computing resources required to achieve their scientific goals. They also voiced the need for a common virtual computational environment that conforms to established standards across the LCFs.
- It would be critically important to identify, apply and follow key interoperability enablers such as metadata conventions and standards, provenance, workflow, data and visualization protocols when developing tools for CESD program and projects.
- The workshop attendees indicated an inventory of available data, compute tools, and resources currently used by CESD and the greater communities are needed. Furthermore, evaluation and assessment of these shared data, tools, and resources would ease the route to adoption into the integrated data ecosystem.
- Workshop participants requested a new class of monitoring services for the next generation of complex workflows. In particular, they want to see services that focus on capturing metrics on data and software downloads, users, and publications resulting from the reuse of their data and software by others.
- LCFs have no policy for retaining data sets with a useful lifespan that extends beyond supported compute facility programs (e.g., the Computational Impact on Theory and Experiment [INCITE] program). In addition, lack of a single sign-on for authentication and federated access was also discussed as a hindrance in using multiple LCF computing hardware and resources.
- Current high-speed reliable data movement is not sufficient for CESD data resiliency and backup needs.
- The CESD workshop attendees understand that in order to be successful they must strengthen their partnership with other national and international agencies.

Additional attendee findings are further elaborated in **Appendix 1**.

2. Background / Introduction

The Climate and Environmental Sciences Division (CESD) within the Office of Biological and Environmental Research (BER) focuses on advancing a robust predictive understanding of Earth's climate and environmental systems by exploiting unique modeling, observational, data, and infrastructure assets, developed and managed by BER. It has a programmatic interest in obtaining systems-level understanding that is driving the need to integrate data and modeling efforts from multiple disciplines. In 2013, the BER Advisory Committee (BERAC) issued a report entitled "BER Virtual Laboratory: Innovative Framework for Biological and Environmental Grand Challenges," which outlined a high-level concept for a potential effort to address the need to integrate data and modeling efforts [BER VL 2015]. The purpose of that report was to assist in the development of a clear vision and potential plans for a federated BER Virtual Laboratory. Such a system would, first, be a unified data construct (e.g., a data infrastructure) where any data can reside and be discoverable, and second, offer a compute environment allowing for rapid model module prototyping, integration, and validation.

Emphasizing data infrastructure needs in the pursuit of exploiting unique modeling and observations, CESD released its own "Strategic Roadmap for Earth System Science Data Integration" report in 2014 [Williams 2014]. The report introduced a data ecosystem that integrates all existing and future distributed CESD data holdings into a seamless and unified environment. The report described a highly coordinated set of data-oriented research activities, with a goal of providing the CESD scientific community with easy

and efficient access to all necessary data archives in order to study increasingly complex scientific challenges. In addition, the report described supporting activities involving metadata compatibility from disparate research projects, fusion of data derived from laboratory studies, field observatories, and model-generated output; server-side analysis; and efficient storage, pattern discovery, and use of DOE Leadership Compute Facilities and networks.

CESD currently supports a variety of observational data archives, including the DOE's National User Facility—Atmospheric Radiation Measurement (ARM) Climate Research Facility's data center [ARM 2015], the Carbon Dioxide Information and Analysis Center (CDIAC) [CDIAC 2015], AMERIFLUX [AmeriFlux 2015], and Next Generation Ecosystem Experiments' arctic, Next Generation Ecosystem Experiments' Tropic, and various terrestrial, and ecosystem science (TES) data archives. CESD also supports the largest model-derived, ensemble-run, data archive used by the international community—the Earth System Grid Federation (ESGF) [ESGF 2015, Cinquini 2014]. In addition to ESGF and the observation-only archives, various test beds associated with observed and various laboratories independently manage model-derived data products such as those associated with the Accelerated Climate Modeling for Energy (ACME) project [ACME 2014].

The CESD data archives and test beds evolved independently of each other to support their corresponding user communities. These archives used domain-specific metadata and data standards for processing, archiving, and distributing their data, and there has historically been little need to focus on metadata compatibility and broader connectivity between their systems and communities. Current research questions of high priority to BER involving complex data from multiple sources (e.g., physical and biogeochemical interactions) are changing the status quo as they require closer collaboration between scientists from different disciplines, and they in turn require better integration of data, tools, and services from CESD and other partner data centers, facilities, and resources [ASCAC Data Report 2013].

To assist in the development of a better-integrated environment, CESD conducted a "Data and Informatics Working Group" workshop to lay the groundwork for a federated BER Virtual Laboratory and CESD's data infrastructure, as described by the previous two reports. For this workshop and report, key CESD personnel (i.e., project leaders, data providers, lead developers, and many others) came together to discuss key cross-cutting requirements. The hope is that this multidisciplinary approach will forge a robust vision for the future in terms of requirements, solution approaches and a prioritized approach to creating the needed capabilities.

Questions addressed at the workshop and in this report include scientific gaps and challenges to be addressed in the planning and development phases of the virtual data laboratory, with emphasis on data infrastructure and the compute environment. Example questions can be found in **Appendix 2**.

This report establishes key community needs and the required deliverables to address these on the basis of clearly articulated use cases from current Subsurface Biogeochemistry Research (SBR) and Terrestrial Ecosystem Science (TES) programs, Environmental Molecular Science Laboratory (EMSL), Regional and Global Climate Modeling (RGCM), Earth System Modelling (ESM), Atmospheric System Research (ASR), Atmospheric Radiation Measurement (ARM) and Integrated Assessment (IA) programs. Furthermore, the use cases were coordinated with appropriate principal investigators of the existing projects and programs.

3. Scientific Challenges and Motivating Use Cases

Any realization of a data and informatics system that serves the needs of BER and its Virtual Laboratory concept must be structured to meet requirements imposed by the science and research activities carried out in support of BER's mission. The particular focus for this workshop was on the data system requirements emerging from multidisciplinary research to understand and predict Earth's climate, its internal variability, and its response to forcing from human activity. There is broad recognition in the

community that modeling, observation, and experimentation are all required to advance our predictive understanding of the Earth system and climate change and that many disciplines and scales of study must be engaged to increase our quantitative knowledge of the Earth as a coupled system.

Figure 1 shows one view of a process of scientific inquiry, hypothesis formation, real-world observation and experimentation, and modeling that serves to increase quantitative knowledge, moving us closer to a robust predictive understanding of Earth's climate and environmental systems. One interpretation of this iterative process is that it encapsulates a large body of scientific and research effort in terms of what we do (blue boxes in **Figure 1**) and what we produce (red boxes). Earth system science is characterized by complexity, in that multiple sets of disciplinary knowledge must be integrated, and interactions among them grasped, with the relevant fields of knowledge and their dominant interactions embracing the spectrum of spatial and temporal scales from cellular to planetary and from fractions of a second to multiple millennia. Data and metadata (descriptive information about the data) are core components at each step in this research, given the inescapable diversity in the types of data gathered and in the ways data are generated, processed, and applied in the pursuit of increased predictive understanding.

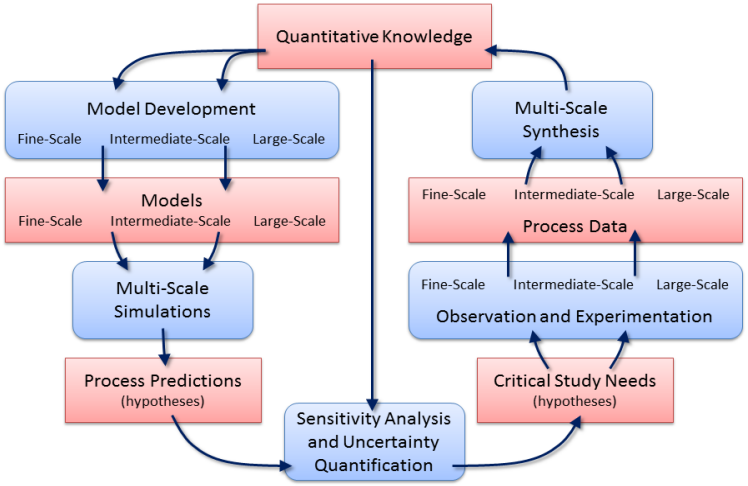


Figure 1. *An iterative approach designed to increase quantitative and predictive knowledge of the climate system through coordinated observation, modeling, experimentation, and uncertainty quantification.*

Rather than attempt an exhaustive assessment of the requirements placed on a data and informatics system by the entire scope of science and research activities relevant to the BER climate mission, the workshop used a small number of use cases to help identify the most significant needs in terms of a system to support what the science community does and what it produces. The use cases presented here are examples. They have enough detail to motivate specific, actionable requirements for a data and informatics system, but they are not intended to cover the entire programmatic scope or range of capability that might be demanded of an operational system. These use cases were designed to represent complex requirements emerging from multi-institutional, multi-disciplinary, and multi-scale investigations, in the hope that cases of this sort would help to identify the broadest outlines of a BER-centric climate data and informatics system and its capabilities.

The use cases themselves are shown in plain text, and specific requirements emerging from individual aspects of the use case are shown as indented *italic text*.

3.1 Use Case #1: Collaborative Scientific Discovery across Discipline Boundaries

This use case illustrates the science requirements placed on a data, informatics, and computation system by a collaborative project involving modelers, computational scientists, data scientists, observationalists, and experimentalists.

A synthesis of previously published observational, experimental, and modeling results has indicated that strong interactions between temperature, humidity, and soil moisture control the composition of

vegetation communities and the fluxes of CO_2 and CH_4 from wetlands in a variety of geographic settings and climate zones.

> *Synthesis studies require indexed access to comprehensive literature and data set resources, with the ability to search and filter on time, geographic location, process-based keywords, investigator, research program, and other fields. Mapping among multiple ontologies or dictionaries is needed to span existing resources.*

A manipulative field experiment is initiated to further explore the influence of long-term warming on CO_2 and CH_4 fluxes in a boreal wetland setting. Pre-treatment observational campaigns have characterized the structure and function of the target wetland.

> *The detailed experimental design should be documented in a searchable format, so that other researchers can find the intended measurements, manipulations, and other background information even before the experiment is running. This should include points of contact for each element of the experimental design so that questions about potential collaboration can be effectively directed and addressed.*
>
> *Pre-treatment observations and characterizations are critical for modeling studies, and should be planned and catalogued with as much forethought and attention to detail, as are the eventual experimental results. Iteration with modeling groups and other experimental efforts is essential to ensure comprehensive pre-treatment observations, since opportunities are necessarily limited once treatments are underway.*

Site-level modeling in advance of the field manipulation indicates that imposed warming will interact strongly with over-winter snowpack, generating a seasonal pattern of positive soil temperature anomalies that are strongest in summer and weakest in winter.

> *A priori modeling is as crucial as pre-treatment measurements to the eventual success and applicability of the collaboration. The bootstrapping nature of this kind of work in a newly developed location means that the data system will need to service a range of synthesis efforts to gather existing driving data, interpolation, and gap-filling methods to make data as relevant to the experimental location as possible, and frequent iteration with the experimental team to ensure that simulations reflect experimental plans. Simulation results need to be made available in searchable format so the experimental team can query potential outcomes and ideally should include the ability to establish what-if scenarios to assess details of the experimental design and measurement plan.*

Additional modeling across a latitudinal gradient of wetland sites indicates a complex relationship among warming, seasonal patterns of soil temperature and moisture, and net changes in greenhouse gas budgets.

> *In addition to intensive modeling at the experimental site, extensive multi-site modeling at other similar locations, potentially including other observational or experimental locations, is a critical step in understanding the relevance of site-level findings. A multi-site simulation capability becomes essential in organizing inputs and outputs as the number of additional sites increases from a few to tens or more. A data and informatics system capable of organizing inputs and outputs and allowing evaluation against a range of observational data types could bring great efficiency to this process.*

As detailed experimental results emerge from the long-term warming experiment, short-term sampling is being carried out across a latitudinal gradient, and both experimental and observational data are being deployed in an uncertainty quantification (UQ) framework to evaluate the model predictions.

> *A system that monitors and reports experimental findings in near-real time can enhance the ability to collaborate with an interdisciplinary and multi-institutional team. Continuity with the*

indexing functions for pre-treatment data, the ability to issue problem reports and updates, and browse-analytics for quick views would all be useful features of such a system.

Field campaign-style sampling across a range of coordinated sites will be a common science requirement in collaborative projects. The ability to replicate a data model with applicability to many sites improves the efficiency of later synthesis and analysis.

Comparison of model results to observations or experimental data imposes a broad set of science requirements. Handling of missing values, unit conversions, spatial and temporal aggregation, scaling of measurement uncertainties in space and time, relative weight assignment to multiple independent observations of the same quantity, and model skill metrics definition are all necessary aspects of a model-data evaluation system. The high dimensionality of model and observational data sets challenges traditional analysis tools and methods, and advance analytics with responsive user interfaces would accelerate new knowledge generation in this area.

At the same time, an effort is underway to integrate representation of wetland thermal hydrology, soil biogeochemistry, and vegetation structure, function, and dynamics into the land component of a coupled Earth system model (ESM). Synthesis studies and multi-site modeling suggest that improved process representation could allow the global model to capture the hypothesized mechanistic controls on wetland carbon cycle and surface energy budgets.

New model development requires a design process that, ideally, is as comprehensive and deliberate as the design of new observational or experimental efforts. A broadly capable data, analytics, and computational system would enable this design process through synthesis and evaluation tools. It would also capture the results of the design process as design documents describing the new process representations, the new model inputs and outputs, and new science requirements for parameterization and evaluation data and analytical resources.

A set of new parameters for the global model must be estimated on the basis of previous literature estimates, new cross-gradient observations, and extensive data collection at the experimental site. A Bayesian UQ framework is deployed to assess model sensitivity to parametric uncertainty, and the most critical parameters are estimated based on multiple independent observational and experimental constraints, each accompanied by uncertainty estimates.

Formal parameter estimation places significant demands on the computational system, as a large number of carefully regulated simulations are typically required. A system that cross-references uncertainty estimates on observational and modeling results is also needed to ensure that empirical constraints are applied appropriately. Analysis of large and multi-dimensional model outputs is required to interpret UQ results. Filtering of sensitivity analysis results produces a reduced set of parameters for formal estimation, but these results can vary in space and time, placing high demands on the analysis framework and requiring engagement of expert knowledge.

The new global model is first evaluated at the site and regional scales against withheld observations and then exercised at the global scale. Global-scale simulations include a series of offline simulations driven with observed surface weather, followed by fully-coupled ESM simulations covering historical and future periods, out to year 2100, under a variety of socio-economic forcing assumptions.

Strict model evaluation using withheld data and/or cross-validation methods provide a conservative estimate of model performance and should be enabled in addition to the more sophisticated Bayesian estimation methods. Challenges here include a diversity of spatial and temporal scales in the available observational and experimental data and the need to aggregate observations or disaggregate model results to make meaningful comparisons.

This is an area of the collaborative use case where existing technologies and practices are already quite mature. Globus, ESGF, and the Coupled Model Intercomparison Project (CMIP)

archive, to name a few, have resulted from long investment in this area. Reproducibility of results and model configurations in the context of large assemblages of sophisticated simulations are also important aspects of the global-scale simulation problem, but they have so far received less attention and have fewer existing solutions.

Single-factor and multi-factor simulations are performed to evaluate the influence of the new wetland model and parameterizations on global scale climate-biogeochemistry feedbacks.

Complex evaluation methods are employed to evaluate system feedbacks and in the areas of signal detection and attribution. Science output and knowledge growth would be enhanced if these complex workflows were incorporated in a broadly capable system.

Results from the global simulations are periodically evaluated against new findings from the experimental site as long-term effects emerge under the warming manipulation.

The overall science objective of hypothesis testing needs to be accommodated in an integrated system. Circling back from global modeling results to evaluate against newly emerging experimental and observational results is a crucial step in that process. A capable system could help in the synthesis of these periodic evaluations, leading ultimately to refined hypotheses and new process investigations.

3.2 Use Case #2: Multi-source Observational Data Integration

This use case highlights the complex relationships that exist among researchers with respect to data collection, stewardship, ownership, and distribution. It also highlights the need for any data and informatics system to maintain transparent records on data provenance and clear guidance on attribution of credit for various stages in the data and project life cycle.

Sally Fields is working at the Harvard Forest and is beginning a nitrogen addition experiment. In laying out her experiment, she has decided what she is going to measure and what metadata she wants to record. She does a quick search for standard templates and metadata specifications and does not find any existing standards for the type of experiment she is doing.

Metadata standards and metadata searching capability with interoperability among multiple data centers at various institutions, agencies, and nations is a core capability that enables all use cases.

Sally begins the experiment and does quality assurance/quality control (QA/QC) on a daily basis as part of her monitoring effort. When she completes the experiment, she does not have time to analyze the data further or write a paper using the experiment results because she has to teach a class. The data are currently stored in an excel spreadsheet. She made up the QA/QC flags she needed to indicate the various situations as they occurred and built a key for the flags as she went along. She also took soil cores, which were analyzed by a commercial lab until that lab went out of business and she had to switch to another lab. She has received the data from the labs. The two labs used different methods to analyze the cores, but both came up with total nitrogen content values for the cores, although one was by weight and the other by volume.

Any metadata standards need to be as comprehensive as possible, but also flexible to accommodate new situations and data types. Quality control information is a necessary component of data and metadata, especially as data products are shared in larger communities where first-hand knowledge of limitations is the exception rather than the rule.

She attends the North American Carbon Program meeting and meets John Flux, who measures similar data in Canada; Dr. Marsh, who has LiDAR data for Harvard Forest; Dr. Nitrogen, who measures leaf level nitrogen at both sites; and Dr. Cycle, who specializes in modeling nitrogen. They would like to work together to do a model validation using the data from the two sites. Dr. Marsh is part of a large data

repository and analysis center, and he offers to host all the data they use for the validation at his site. Sally is a bit worried about this, since she does not want the data to be available outside this collaboration until she writes a paper about her results for her tenure case.

> *Productive collaborations require recognition of diversity in requirements and expectations for data sharing, and a broadly capable system needs to both record and protect the interests of multiple parties.*

After they get started, Dr. Flux decides that he does not have the time to contribute to the writing of the paper but that the group is still welcome to use his data as long as he receives credit for it; he has not yet written a paper based on the data and is a bit worried about continued funding for data collection. (More generally, he is making his data available to any interested user via Dr. Marsh's system but would like credit for his contribution.) He is also concerned that there might be QA/QC problems with the data and would strongly prefer to see any results based on the data before they are published. Dr. Nitrogen has already provided his data to AmeriFlux and it is available there with a digital object identifier (DOI), so he asks the team to use that version and cite the DOI. Dr. Cycle has not yet published a paper about his model and is not yet ready to release the model, so she would prefer to run all of the validations on her own cluster.

> *The lifecycle of data and information in a multi-partner collaboration can be complicated, and provenance information that shows previous, current, and planned future stages in the lifecycle for a given data set needs to be maintained and amended as the collaboration proceeds.*

3.3 Use Case #3: Climate Modeling and Model Analysis

This is a pair of use cases that deal specifically with generation and analysis of large volumes of data from single or multiple ESMs running one or more simulation experiments, sometimes with multiple ensemble members per experiment. Different types of analysis invoke different data storage, handling, and processing requirements.

3A: Model intercomparison for study of extra-tropical cyclones. Dr. Bigdata is conducting a study of extra-tropical cyclones (ETCs) and how the frequency and severity of ETCs will change under different future carbon scenarios. The source data set is the CMIP, phase 5 (CMIP5) model data. The time scales necessary for tracking evolving weather systems are relatively short, so six-hour data sampling is required, which results in an initial data set much larger than those assembled by most scientists (many tens of TB). Only a small subset of the data is actually needed for the analysis, but there is no canned analysis tool available for this purpose at the globally distributed data centers that host the CMIP5 archive.

> *While many routine search and subsetting capabilities are supported by the existing network of climate model data centers, new and innovative analyses are constantly emerging. These can place unforeseen demands on the existing data systems, meaning that a flexible and configurable capability must exist in addition to the standard hosting facilities. This may take the form of a prototyping environment, but the storage and processing requirements for new prototype analyses can be very large.*

Dr. Bigdata identifies the model variables required for the analysis and submits a query to the ESGF data infrastructure. This request results in a set of Globus data transfer jobs that deliver the data from the ESGF data infrastructure to a file system at an ASCR computing center. Once the data arrives on the file system, Dr. Bigdata then runs a secondary processing code that requires the massively parallel environment of a national computing facility. The result of this secondary code is a high-value data set of significantly reduced scale, which provides significant leverage for all downstream analyses, especially if other scientists can publish the derived data set with metadata that facilitates interpretation of the data.

The prototyping and secondary analysis environment needs to be close to high-throughput data transfer networks and needs access to high-end computational power. The ESGF data infrastructure must also be able to support the necessary large-scale data transfers to the computing facility, which has sufficient capability to run the analysis. New value-added data products need to enter a data lifecycle tracking system that ensures proper metadata, indexing, and attribution is generated and retained.

Dr. Bigdata's colleague, Professor Sandy Katrina, is studying the effects of climate change on tropical cyclones. Instead of obtaining data from a distributed CMIP5 data set, she creates a modestly large set of multi-decadal simulations by running the ACME climate model at a 25 km horizontal resolution. Upon completion of her simulations, she runs the same secondary processing code to identify storms and their tracks. She then uses that track data to query a large 3D sub-daily data set to examine changes in storm structures.

Many climate modeling applications require dedicated analysis and data-reduction capability at the high-performance compute sites where models are run. In addition to compute capacity for data reduction and batch-mode analysis, interactive visualization capabilities can accelerate the identification and extraction of new knowledge from large multi-variate data sets.

3B: Three-dimensional ocean analysis. Dr. Lotte Malte-Modele is studying the global ocean and how oceanic variability and change are evolving due to a series of future CO_2 scenarios. The source data set is from CMIP5 and CMIP, phase6 (CMIP6) models. Due to local storage limitations, Dr. Malte-Modele would like to undertake a considerable data reduction on the ESGF nodes where the data resides, thereby reducing the total local footprint required to store the analyzed outputs and decreasing data transfer volumes. As part of the data reduction, Dr. Malte-Modele needs to analyze data on the native grids provided by the modeling centers, often performing calculations that require careful treatment of computed transport. For this, she needs specialized software that is "grid aware" and considers cell volume weights during calculations.

To ensure scientific validity and publishable results, analysis software must meet exacting technical requirements. General-purpose software may not be fit for special-purpose analyses, and a data and informatics system needs to be explicit about the capabilities and limitations of default software, while accommodating special-purpose software.

Dr. Malte-Modele identifies the ocean variables required for the analysis, and submits a query to the ESGF archive to obtain a list of all available data located across the federated archive. She then constructs an analysis script using the UV-CDAT analysis package, which is co-located with the data on each ESGF node. Thanks to local resources available on ESGF nodes, this task is completed within a couple of hours. These reduced data are then transferred to local storage, using a series of Globus data transfer jobs initialized at one of the ASCR computing centers. Using local software stacks, Dr. Malte-Modele then undertakes the final stages of her research using an array of analysis and graphics tools to prepare publication-ready figures.

Some challenging data analysis and handling requirements are already met by existing systems, and so a BER-centric effort for data and analytics will not need to start from a blank slate.

4. Survey Results

As preparation for the workshop, the organizing committee carried out an online survey of DOE BER CESD scientists. The survey aimed to ascertain what they felt were the greatest needs for additional support. The request for feedback was sent not only to the workshop participants, but also to BER CESD PIs. For the majority of the questions, the users were presented with a scale from one to

The wider CESD community identified knowledge capture, management, and sharing as a key development area.

six to indicate if they saw a need for additional support (one indicated no or little interest, while six indicated a highly important area for development). The survey calculated both average values for each question across all responses (i.e., a value of 4.79 would indicate that most responders would rate this topic as of high or very high interest), as well as percentages of responders that gave this topic a particular rank (i.e., 41% ranked this as very high).

The survey asked the responders to self-identify as data providers, resource providers, software developers, climate modelers, or data analysts (see **Table 1**).

Table 1. *Self-identification categories for the 75 scientists who responded to the survey request.*

Scientific Background	Description	Total
Climate modeler	One who develops the quantitative methods to simulate the interactions of the important drivers of the Earth's climate such as atmosphere, oceans, land, and sea ice.	15
Climate model data analysts	Analyze output to understand simulation and observational output for knowledge discovery and change.	18
Resource provider	Technology provider of hardware and software resources at high-performance computing facilities.	4
Software developer	A person who develops stand-alone software for the climate community. Also known as computer programmer, application developer, and system software developer.	6
Data provider	The person responsible for providing data and metadata (describing the data) to the community. Also responsible for the quality of the data. Associated with climate modeling groups and associated data center.	32
Total		**75**

Forty-percent of responding scientists saw access to sufficient computational and storage resources as a very high need. Also notable was the emphasis placed on more reliability and resiliency in the resources and services provided to them (34% identified this as their highest need) and access to sufficient observational and experimental capabilities (26%). Data and software resources were identified as the most difficult to discover with 40% of respondents stating that they might need hours or days to find what they needed. Matching with this were requirements for more user support for data access and usage (30%), data publishing (26%), and data sharing (23%). Of relevance to our efforts to design a more integrated data and computing infrastructure was the finding that the large majority of scientists access data and compute resources via web interfaces or remote login rather than application program interfaces (APIs) and are therefore currently not set up to flexibly leverage more integrated capabilities across the DOE complex.

Table 2. *Top ten needs identified by the survey.*

Survey Question	Average Rating or Percentage in Highest Need Category
An easy way to publish and archive data using one of the DOE data centers	4.79
A means for comparison of diverse data types generated from observation and simulation	4.71
User support for data access and usage	4.64
Access to sufficient observational and experimental resources	4.58
Access to enough computational and storage resources	4.52 / 41%
Method of ingesting and accessing large volumes of scientific data (i.e., from data	4.49 / 39%

Survey Question	Average Rating or Percentage in Highest Need Category
archive to supercomputer)	
Quality control algorithms for data	4.46 / 31%
A unified and single user account to access all BER and ASCR resource	4.44 / 38%
Reliability and resilience of resources	34%
In-situ analysis of observational, experimental and computational results: the ability to interpret results and verify new insights within the context of existing scientific knowledge	4.4

The survey questions were divided in to different categories; out of these knowledge gathering, managing, and sharing (KD) was identified as the overall area of greatest need followed by human and computer interaction (HCI). The topics covered in these categories can be found in Table 3.

Table 3. *Top needs identified by survey respondents.*

Survey Questions	Average Rating
KD- Method of ingesting and accessing large volumes of scientific data (i.e., from data archive to supercomputer)	4.49
KD- Quality control algorithms for data	4.46
KD- Interfaces that ensure a high degree of interoperability at format and semantic level between repositories and applications	4.18
KD- Provenance capture information for data	4.11
KD- Reproducibility	4.06
HCI- Collaborative environments	4.31
HCI- Improved user interface design	4.00

At the same time, questions focused on the effective use of exascale systems received mixed results, pointing to a potential need for further education of the wider community. For example, new techniques for working with deep memory hierarchies on extreme scale computing systems reached only an average of 3.24, and the direct data delivery into ASCR computing systems from BER data resources faired only marginally better with a score of 3.86. On the other hand, ingestion and access to large volumes of scientific data garnered a score of 4.49/39%, and the petascale-related topic of in-situ analysis of observational, experimental, and computational results achieved 4.40.

A list of questions ranked by average rating can be found in **Appendix 3**.

5. Data Services Needed to Support Science Requirements

After discussing and developing use cases that exemplify the scientific goals and challenges that the climate research community face today, a workshop session focused on exploring

The CESD community identified requirements for additional enabling data capabilities throughout the full research lifecycle from data discovery, multi-source data treatment handling large data volumes, flexible data analysis tools, reproducibility, and data publication and attribution.

these in more detail was held. In particular, the participants were asked to identify the key data and computing challenges that the community encounters and the types of services that would have the most impact on their scientific discovery process. Workshop participants were split into two teams to discuss this topic, but the responses from both teams were similar in content.

Researchers identified the following data-related challenges in their research processes:

- *Data and Linked Resource Discovery:* Workshop participants identified the time-consuming search for older data, potentially from papers, as an obstacle. Further, they noted that data discovery alone was not sufficient, as researchers also needed to be able to identify related metadata, provenance, and tools to use the data with confidence and ease. Similarly, they needed discovery methods with suitable computational and storage resources to analyze any identified data.
- *Multi-Source Data Treatment:* Integration, correlation, and comparative analysis of data with different dimensionalities, geophysical properties, levels of data quality, and related uncertainties are domains where few solutions exist. A particular challenge is the comparative analysis of observational and modeling data. There is a perceived lack of dialogue about data harmonization between the two communities. Several workshop participants stated that in recent years efforts have been made by DOE climate programs (i.e., ARM, ASR, RGCM, and ESM) and others around the world to improve connections between the two communities. More recently, CESD's ESS conducted a workshop on model-observation integration, modeling framework, data management, and scientific workflows [ESS Report 2015].
- *Handling Large Data Volumes:* Analytical tasks use increasingly large data volumes from multiple geographically distributed resources. The community is looking for new approaches that enable the efficient analysis of these data sets without the need for massive data transfers.
- *Flexibility of Tools:* Workshop participants identified a range of data tool challenges such as the ease of adoption, scalability, and adaptability; determination of future needs; and issues with cross-tool integration and accompanying training and education. Many useful community tools were developed in a different era, when data volumes were smaller and analysis processes were carried out on local systems with single processors. The community would like to continue to utilize the knowledge and capabilities encapsulated in these tools (e.g. trusted, community-wide, standardized mathematical approaches), but do so in a more scalable environment with more modern user interfaces. Advice was also sought in terms of good data models and approaches that will make it easier to integrate tools in complex data analysis workflows.
- *Reproducibility:* Scientists are looking for practical solutions to enable reproducibility of their work, be it modeling or data analysis tasks. Their focus is primarily scientific reproducibility (replication of conclusions with different methods) and computational reproducibility (the same results with the same modeling setup).
- *Data Publication and Attribution:* Researchers are looking for guidance and support on standardized ways to publish data that integrate well with the community's journals and their expectations. Furthermore, provisions need to be made to ensure that all researchers have access to the required long-term storage and curation capabilities that would accompany these formal data publication efforts. A central discussion point was attribution, which must go hand in hand with the data publication effort. Data products are often based on the work of many others. Data sets are integrated and refined as they through different phases of the research process, from raw data collected from heterogeneous data products to the final publication of a data set used for the validation of a climate modeling campaign. Community-determined standards are needed regarding who should be cited at which step. Also discussed was the concern that data could be used inappropriately so researchers are seeking methods to engage with others on subsequent use of their data.

Based on the identified challenges, the two teams discussed what general data services would be needed to address their challenges. Solutions included:

- Publishing a notional data service architecture—a taxonomy of what data services are provided and where.
- Partitioned data service by size (downloadable vs. very large).
- Discovery based on metadata that describes the conditions/context under which data were collected.
- Standardized data and metadata formats across observational and modeling data to enable easier integration and comparison.
- Server-side computations that push algorithms to the data rather than downloading the data.
- Intelligent data services that inform the user of other related data products that may interest them (data recommendation engines).
- Capability for providing a persistent link to a specific data group that can be published and/or accessed in the future without repeating a complex search.
- Means of avoiding duplication of data downloads to community computing resources.
- Providing programmatic access to data services so that these could be easily used in scientific workflows.
- Provision of collaborative workspaces.

The four highest priority requirements identified are: server-side data subsetting and analysis; better data documentation; sufficient data and computing capacity, including dedicated resources for data science; and standardized interfaces between tools and infrastructure services.

Feedback indicated that scientists would like to find synthesized observational data products (ARMBE, OBS4MIPs, etc.) in support of model development and evaluation in such a data environment. These data sets should also be accompanied by robust QC algorithms, UQ assessments, and linked to tools that support the merging, processing, and further analysis of the data. Furthermore, they should be easily accessible and usable in model development test beds.

Research communities and data service providers see two key impediments to creating these types of services: lack of dialogue and coordination across disciplinary boundaries, and lack of funding for such efforts—i.e., stabilizing the funding stream for long-term operations. Should those key impediments be addressed, the software developers highlighted a number of additional challenges such an effort would face. These included overcoming current requirements for multiple authentication and authorization layers, making sufficient computational resources available, and creating sufficient data and scientific expertise to enable all to participate in this new environment.

Scientists agreed that a successfully implemented infrastructure would not only speed up the scientific discovery processes through higher performance tools and removal of redundancies but also more importantly, enable new science and discoveries through easy experimentation with novel data analysis approaches.

Attendees of the workshop voiced strong concerns regarding the lack of storage and computing resources required to achieve their scientific goals. They also voiced the need for a common virtual computational environment that conforms to established standards across the LCFs.

6. Advanced Computational Environments and Data Analytics

Advanced computational environments, supported by key climate modeling, observations, and data centers such as those hosted at DOE's NERSC and the Argonne and Oak Ridge Leadership Computing Facilities, provide the CESD community with high-performance computing (HPC), clusters, robust short- and long-

term storage, networking, and coordinated software resources and tools. These major computing facilities currently have different architectures (e.g., graphics processing units [GPUs] vs. accelerators), programming models, and operating environments (i.e., hardware, software, policies, security layers, queue management, etc.) running on multiple systems. In addition to running and processing state-of-the-science climate information at these DOE compute facilities, the community must rely on multiple levels of services to effectively manage, analyze, and visualize distributed data from many sources.

Moving from the present computational environment to a federated system of tools and services will require, among other tasks, ensuring that the following levels of services (visible in **Figure 2**) be robust, resilient, and consistent throughout:

- *Common Data Services:* Shared across all CESD projects and hopefully with other sciences as well, such as movement, curation, and long-term preservation, discovery, exploration, etc.
- *Domain-Specific Distributed Data and Analytical Services:* Captures the set of unique requirements and needed services for each unique CESD climate project. For example, software performance (i.e., parallel input/output [I/O], analysis, and data set transformation) and data analysis services with better I/O bandwidth and more memory for analyzing and computing ever-expanding data sets.
- *Data Systems Software Layers:* Includes standardized lower layers of software services such as metadata, directory structures, provenance, extending bit-level verification and workflows that allow reliable and unlimited access to computational and analytical resources with well-defined, scriptable community APIs. Another avenue of services includes the ability to reliably archive and serve data where the user can adjust the cost, speed and reliability of the underlying storage service.
- *Data System Computational and Storage Hardware:* Includes HPCs, clusters, clouds, and dedicated large-scale archives, for modeling, in-situ data analysis, and post-hoc large-scale computational data analysis. This also includes in-transit processing to enable extreme-scale climate analysis. There is also an emerging ability to provide high reliability, geographically distributed storage which should be further explored.
- *Networks:* Binds the collection of disparate hardware, other networks, and software resources for CESD community use. Networks are also necessary to replicate and move large data holdings at storage facilities and to federate connectivity. ESnet's 100 gigabit (Gb) network is of particular interest, along with facility implementation of data transfer nodes. Connections between the facilities and the community imply improvements to Globus/GridFTP and data endpoints (i.e., disk-to-disk, disk-to-tape, etc.).
- *Portability:* Operating environments and methods between flagship computing facilities must not be unique and allow science flows between the centers to be interchangeable. ACME is one example where workflows must reliably operate the same across the Leadership Computing Facilities (LCFs).
- *Support:* User support for reliable access to computational resources, data transfers, login access, persistence data preservation, stakeholder training and outreach, and general system use and documentation.

If CESD is to optimize its investments in data, and therefore the scientific impact of its observational and modeling programs, it must ensure that a common virtual computational environment is in place and a significant fraction of that environment is shared among the different CESD and international community activities, rather than having specific domain environments for each project. Therefore a comprehensive, long-term, sustainable solution for empowering domain-specific distributed data services, data system software layers, next-generation HPC and storage, and next-generation networks accessing national and international large-scale data sets must be an integral part of CESD's overall science strategy. Community-established standards and protocols are needed for distributed data and service

interoperability of independently developed data systems and services. A reference model and supporting API standards are essential for enabling collaborations and facilitating extensibility whereby similar, customized services can be developed across CESD science projects, as shown in **Figure 2**. The environment must support the ability for resources contained at every level of the figure to transfer information within and across the multiple layers of services.

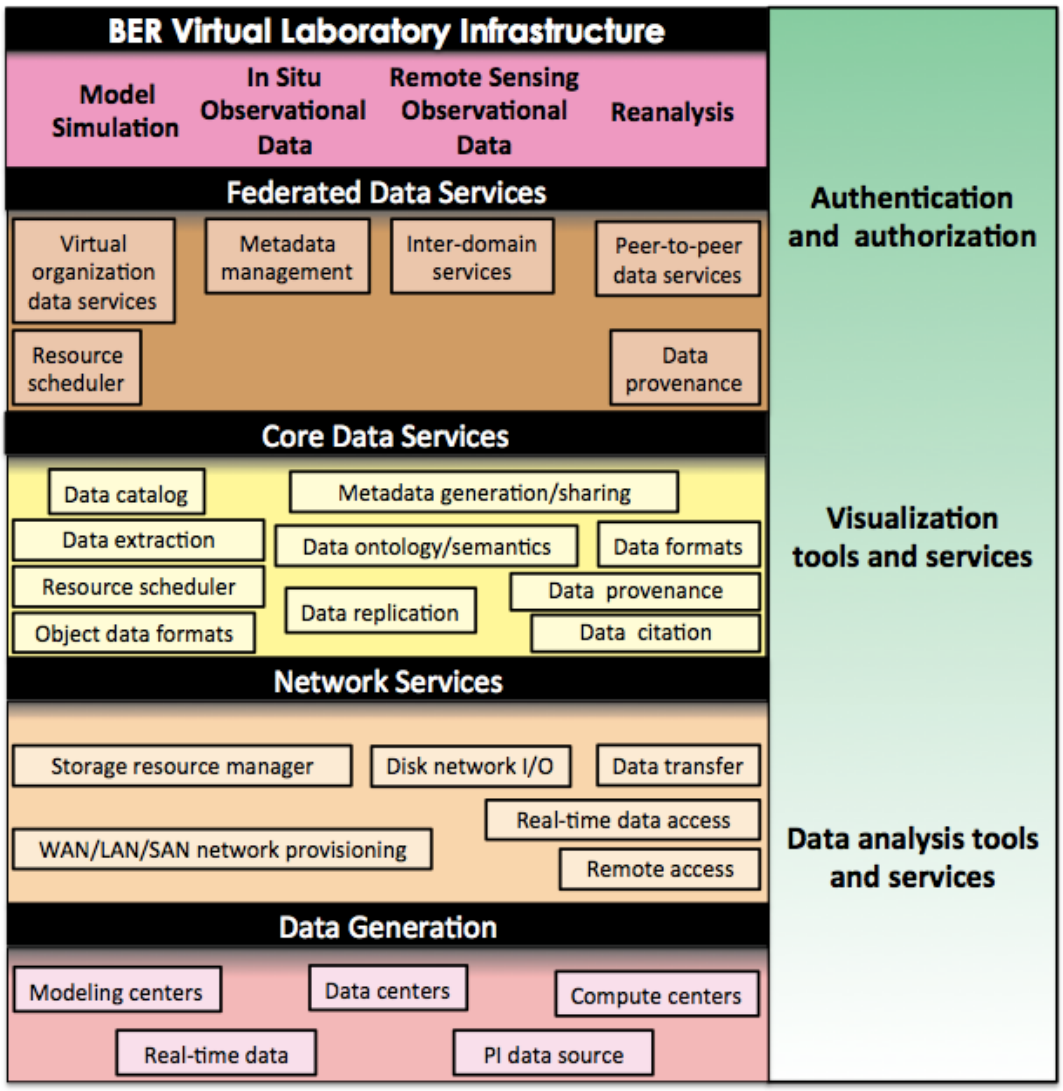


Figure 2. *Framework and relationships for distributed federated climate data products and services in order to support a powerful and flexible advanced computational virtual environments and data analytics. Each service hosted will be exposed through a set of simple and well-documented Web-service APIs—layered with security when appropriate—so that clients of different kinds can easily execute invocations and perhaps chain requests in complex scientific workflows.*

To address usability issues, more comprehensive and constantly up-to-date documentation would exist to aid scientists in hardware, software, and infrastructure discoverability, availability, and access. Key hardware issues include storage, cores, memory, and compute interactions. Today, the use of hardware has a steep learning curve, with multiple levels of integral security details (such as credentials, authentication/authorization, tokens, VPNs, etc.) and each compute facility restricting resource and service use. Managing and analyzing distributed data for petabyte archives consisting of 100-terabyte data sets necessitate both long-term storage for observations and short-term scratch space for large-scale

computational experiments. Diversity of compute resources must be standardized across the facilities such that similar programming models (such as FLOPs-intensive vs. data-reducing) are reliable, resilient, and above all, consistent among the virtual facilities. Containerized performance-portable methodologies could be addressed by multi-level computing approaches with shared storage and archival high-end compute-intensive, mid-range data-intensive architecture, and typical cluster resources. This will also include compatible I/O and memory performance for large-scale data sets. Usability of the system should not exclude non-expert users from accomplishing large-scale data analysis and should allow all users simple navigation of the batch queuing system.

If data are housed at a major facility or data center or distributed across many, it is feasible to move large amounts of that data in a reasonable amount of time to compute facilities (for remote processing), or to data storage (for replication and backup). This will allow federation of data to be managed differently than the way users interface with it today (i.e., most users download data to their home organization for performing analysis and visualization). Once data has been created, produced, or reduced, there is a need to publish or republish data as a service so that it is usable by other members of the community without large-scale data movement. In this way, remote or local data manipulation and publication can be made available to all, including cloud services that will complete the full spectrum of data availability and accessibility.

From the resource providers and software developers' perspective, the primary impediment to computational environment and data analytics development is continuity of funding. Keeping up with heavy user demands and disruptive technologies for this type of environment will require sustained monetary resources. Therefore, a sustainable business model for CESD-wide data infrastructure and environments is warranted; cost justifications and metrics of success will be evaluated and determined in terms of scientific productivity enhancements. Additional key impediments include remote compute services and more short- and long-term storage (i.e., rotating and tape archives).

The prioritized needs for the virtual federated computational environment include:

1. *Hardware:* More storage (petabyte-scale), compute cores, and memory for co-located data computing. Coordination of hardware efforts with ASCR petascale and exascale HPCs is also a must. Especially noted during the workshop are difficulties with compute core technologies compatibility with the software; indicating that ever-changing code revisions are needed as the technology shifts back and forth.
2. *Simulation and observational storage and preservation strategy:* Publishing data so that it is usable by other members of the community. Preliminary inference from the use cases indicates 500 TB of data a year for average CESD project publication. One or two CESD projects (such as the CMIP) expect to publish tens of petabytes of data over the lifespan of the project.
3. *Data analysis, retrieval, and reduction:* Standardization on analysis framework. Fat nodes with high-throughput input/output and memory.
4. *Support:* Computational and analysis classes on the federated environment.
5. *Documentation*: Up-to-date documentation detailing resource availability and specific up-to-date user guides for analysis packages. It would also be useful to provide users with training and access to white papers that outline next-generation computational environments so that DOE science and CESD infrastructure can evolve in lock step and upcoming projects can make the most use of new available resources.
6. *Operational support*: Facility support for operational services and data archives (e.g. CMIP).

The virtual federated environment must also allow scientists to access and compare observational data sets from multiple sources including, for example, the Earth Observing System (EOS) satellites and the ARM sites. These observations, often collected and made available in real time or near real time, are typically stored in different formats and post-processed to be converted to a format that allows easy comparison with model output (i.e., CMIP). The need for providing data products on demand, as well as

value-added products, adds another dimension to the needed capabilities. Finally, science results must be applied at multiple scales (global, regional, and local) and made available to different communities (scientists, policy makers, instructors, farmers, and industry). However, providing results to the science community will take precedence over all other user communities.

7. Data Centers and Interoperable Services

DOE-supported data centers handle diverse scientific data products, from multiple petabytes of climate model data to field and experimental data. These data centers use a variety of tools and technologies to manage and share their data. Some data centers also provide interoperable data and services to broader scientific communities (for example, Obs4MIPs, THREDDS data catalog, and ISO-19915 metadata standards). **Figure 3** provides a concept diagram of an integrated cyber-infrastructure using various interoperable services.

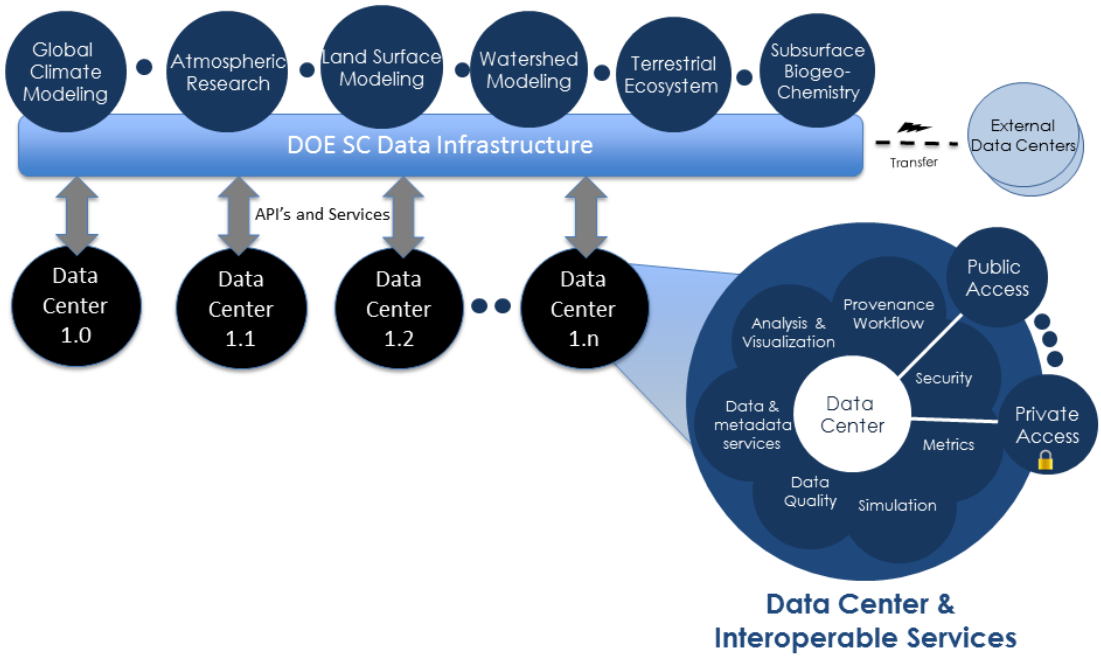


Figure 3. *A concept diagram from the BER Data Strategic Roadmap document for an integrated cyber-infrastructure leveraging core Office of Science resources to enable discovery, analytics, simulation, and knowledge innovation.*

During this breakout session, the participants discussed the following BER CESD and other data centers and their current interoperable services. Key points from this breakout discussion follow.

> It would be critically important to identify, apply and follow key interoperability enablers such as metadata conventions and standards, provenance, workflow, data and visualization protocols when developing tools for CESD program and projects.

a. Earth System Grid Federation (ESGF)

The Earth System Grid Federation (ESGF), one of the largest-ever collaborative data efforts in climate science, is now used to disseminate model, observational, and reanalysis data for research assessments and model validation (see **Figure 4**). ESGF is an international multi-agency driven activity led by DOE as an open-source, operational code base with secure, petabyte-level data storage and dissemination of the resources essential for studying climate change on a global scale. ESGF is designed to remain

robust even as data volumes grow exponentially. Virtually all climate science researchers in the world use it to discover, access, and compute data. The decentralized approach to ESGF has changed relatively recently from a client-server model to a more robust peer-to-peer approach already proven for distributing large amounts of data and information. A system of geographically distributed peer nodes comprises ESGF. These nodes are independently administered yet united by common protocols and interfaces, allowing access to global atmospheric, land, ocean, and sea-ice data generated by satellite and in-situ observations and complex computer simulations for use in national and international assessment reports. Scientists are accessing climate data more efficiently and robustly through newly developed user interfaces, distributed or local search protocols, federated security, server-side analysis tools, and other community standards—all for improving our understanding of climate change.

ESGF's architecture can easily be leveraged for accessing data from other scientific domains, such as satellite, instrument, and other forms of observation data. ESGF is now in the early stages of being adapted for use with NASA DAACs, NOAA's National Centers for Environmental Information published data archives, and the international communities' data exchanges. The importance of ESGF continues to grow as computing platforms and archives expand and reach extraordinary speeds and capacity.

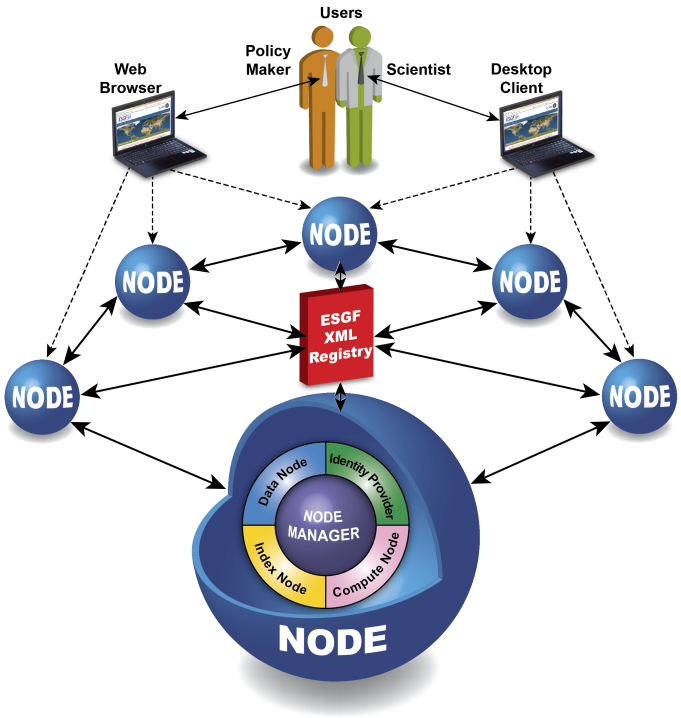


Figure 4. *ESGF ensures equal access to large disparate data sets (i.e., simulation, observation, and reanalysis), which in the past would have been accessible across the climate science community only with great difficulty. The ESGF infrastructure enables scientists to evaluate models, understand their differences, and explore the impacts of climate change through a common interface, regardless of the location of the data.*

b. ARM Data Center

The ARM Climate Research Facility operates field research sites around the world for global change research. Three primary locations—the Southern Great Plains mega-site, North Slope of Alaska mega-site and Eastern North Atlantic site in the Azores—plus aircraft and the portable ARM Mobile Facilities are heavily instrumented to collect massive amounts of atmospheric data. As part of this effort, ARM scientists and infrastructure staff provide value-added processing to the data files to create new data streams called value-added products. In addition, the ARM Data Center archives and distributes PI-contributed and field campaign data products.

The ARM Adaptive Architecture (**Figure 5**) is being developed to provide the data tools, connections, and software for scalable micro-services to support diverse observational data sets. Many interoperable services such as machine-readable data quality, data flow monitoring, next-generation data discovery, visualization, data extraction, and analysis capabilities will be delivered through tools such as the ARM Data Integrator (ADI), Python ARM Radar Toolkit (Py-ART), Data System Status Viewer, Data Delivery Tracking, PI Data product registration (OME), Data Discovery portal, Data Citation tool using automated DOI generations, THREDDS, Big Data analytics using No-SQL (Cassendra and Hadoop), and data visualization tools.

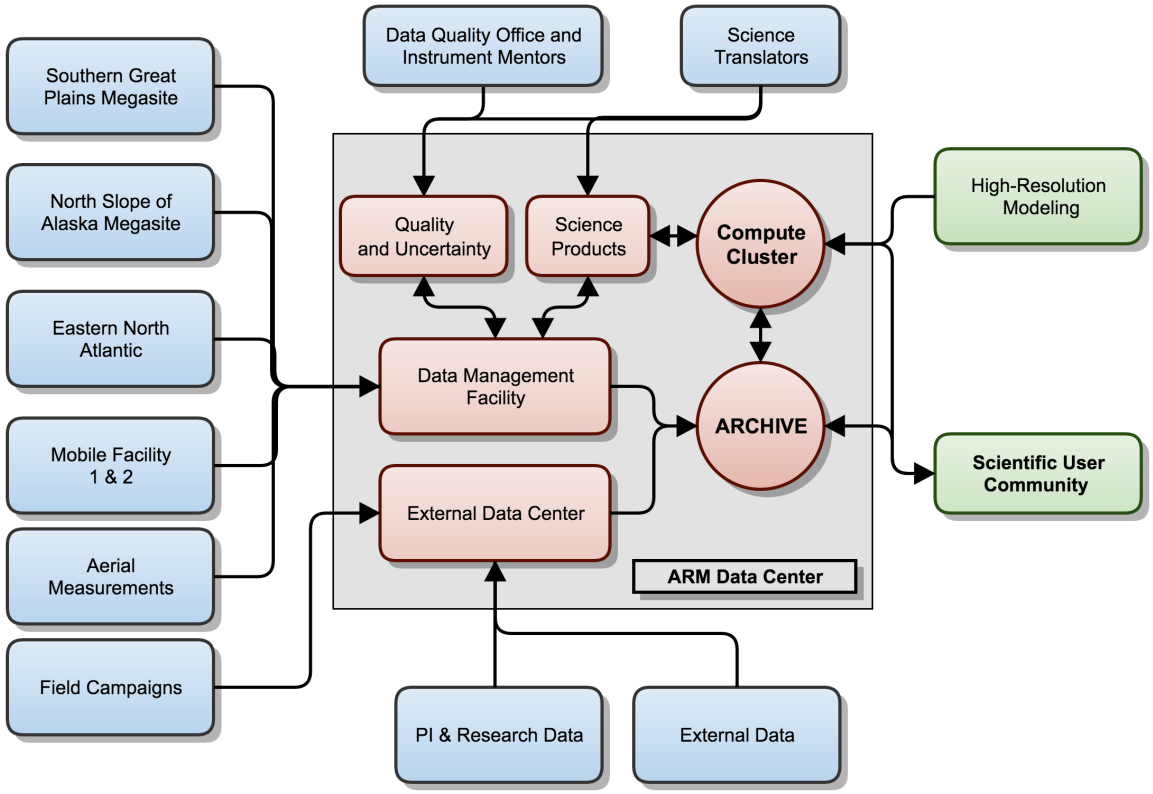


Figure 5. *The ARM Adaptive Architecture is being specified and evolved to provide the data tools, connections, and software for scalable micro-services.*

c. Carbon Dioxide Information Analysis Center

The Carbon Dioxide Information Analysis Center (CDIAC) offers publicly available data and value-added products for climate change research. CDIAC's data collection is diverse, reflecting the breadth of climate change research, and includes atmospheric, oceanic, terrestrial, climatic, and anthropogenic emissions holdings. Considerable effort is devoted to the development and production of science-driven global and regional scale synthesis products (e.g., AmeriFlux, GLODAP, global and national fossil-fuel CO_2 emission estimates). CDIAC hosts and serves processed data from measurement networks (e.g., AGAGE, TCCON), intensive field campaigns (e.g., NGEE Arctic, SPRUCE, HIPPO), and projects (e.g., Global Carbon Project). A searchable catalog based on standards-compliant metadata enables easy data discovery and customized interfaces allow users to query, visualize, subset, and download many CDIAC collections. Multiple data formats are offered for most data holdings to facilitate broad use.

CDIAC is evolving from an independent data center to an integral part of a federated data system that includes the ESGF, ADC, and NASA DAACs (**Figure 6**). As part of this federated system, CDIAC will develop data tools and services to facilitate interdisciplinary research across multiple data holdings and scales and will benefit from existing tools and future developments elsewhere. Existing workflows, processing capabilities, and automation will be leveraged and expanded to support existing and future DOE ESS research.

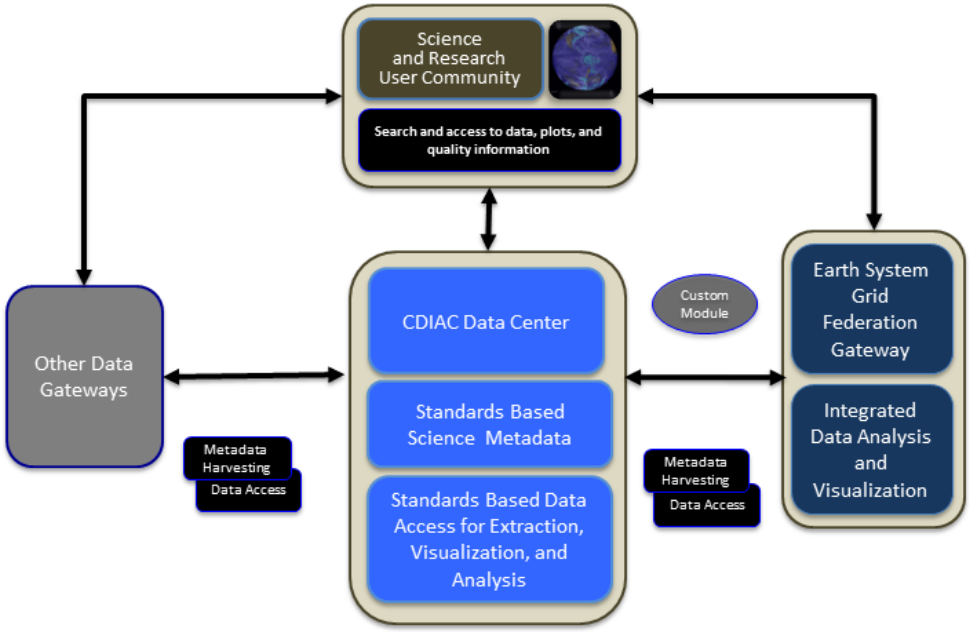


Figure 6. *Diagram illustrating publication of CDIAC data and metadata to the Earth System Grid Federation.*

d. Other Interoperable Services

Other BER data centers, such as EMSL and their interoperable services will be included in future discussions. Globus (www.globus.org) provides high-performance, secure, and reliable data transfer, sharing, synchronization, and publication services for the science community; in addition to user-friendly Web interfaces and simple APIs make it easy to integrate them into services such as ESGF, NCAR's Research Data Archive, and the DOE BER KBase. The ORNL CADES infrastructure offers hardware hosting with the ability to deploy custom software stacks to meet diverse user needs and should also be considered. In addition, workshop participants also discussed various data services from external data centers such as NASA's Giovanni re-analysis and re-gridding service, OGC and metadata services offered by NASA's Distributed Archive Centers (10 data centers), and satellite data services offered by the NOAA's NCEI (formerly NCDC) snow and ice center. In addition the group also discussed about the mega-portal services offered by NSF's Unidata/UCAR/NCAR data stewardship engineering team.

The workshop participants prepared the following list of required and recommended interoperable services to communicate between centers.

- Server-side analysis and visualization (analysis-as-a-service vs. downloads) should be scalable, robust, resilient, easy, tested. These services should allow users to cache recent analysis in a sharable way (re-use analysis where possible) and isolate function from implementation (analysis should specify what is done, not on which machines [i.e., say "no" to shell scripts]).
- Common metadata across data sets should be based on properties/features, temporal-geospatial, variables, and should include provenance and versioning for reproducibility (**Figure 2**).
- Seamless unified search and access across data sets is a critical component for enabling interoperability. These search capabilities should provide common indices, hashes, duplicate detection, and quality control info/method/data where possible. It should allow API-based access to data sets, data services and catalogs, measured reuse of data, citation, acknowledgments etc.
- The data services are preferably built based on open-source software licensing. The team also suggested that curation and stewardship policy comparison across centers should be considered;

this includes: index of policies in forming DMPs, data persistence, number of copies, and policy best practices. In addition, high-speed data transfer services to support large-scale data analyses built using current best practices such as large-scale data movement infrastructure using the Science Demilitarized Zone (DMZ) model [ScienceDMZ 2015] and Globus services [Chard et al. 2014] should be considered in enabling the interoperable services.

8. Inventory of Existing CESD Computer Resources, Data Tools, and Services

CESD data projects use a variety of data management tools and technologies, many of which are open source based, community developed, and used by multiple projects. The data tool needs of the projects are diverse and span a wide array of capability needs. There is not currently one tool able to handle all of CESD's diverse data needs. Nor, despite the wide array of tools, are all the needs being presently met; gaps still exist in such areas as QA/QC, interaction with gridded data, and metrics.

> The workshop attendees indicated an inventory of available data, compute tools, and resources currently used by CESD and the greater communities are needed. Furthermore, evaluation and assessment of these shared data, tools, and resources would ease the route to adoption into the integrated data ecosystem.

The desired goal is a healthy and sustainable ecosystem of tools that together serve the diverse data needs across the projects. The first step toward meeting that goal is to develop an inventory of existing data management tools used within CESD projects (see **Table 3)**. Next, benchmark testing of existing tools needs to be carried out to evaluate their potential for broad adoption within the virtual laboratory infrastructure. In addition, work on standardization of storage formats, APIs, authentication/authorization, and identifiers can significantly improve interoperability between tools and enable a healthy competition between tools available without compromising interoperability.

Table 4. *Open-source tools that should gain wider accessibility within the CESD community.*

Tool	Need
Infrastructure	Use flexible, extensible infrastructure tools for future CESD efforts and partnering DOE projects to automate laborious, repetitive simulation data tasks and heighten productivity and user experience. The same infrastructure must allow CESD scientists to access and compare data sets from multiple sources (i.e., simulation, reanalysis, and observational satellites and instruments).
Globus transfer, sharing, publication; GridFTP	
PerfSONAR	
Panda Global data and job placement across facilities	
ESGF	
Velo	
Docker	
PerfSONAR for network data transformation	
ARM Data Integrator	
Py-ART	
Serial/parallel tools: NCL, NCO, CDO, UV-CDAT (need scalable versions), multipurpose	
CMOR (generates and checks for CF standards)	
ILAMB, ESMF regridder	
TECA and Illiad for analyzing HDF5 atmospheric data (high performance)	
Metadata	Metadata tools to discover, facilitate, and navigate the CESD data infrastructure.
Online Metadata Editor	
Mercury, ES-doc	
OpenDAP, THREDDS	
User metrics and usage analysis	Service-specific metrics to measure the usage and adoption of specific capabilities.

Tool	Need
Data quality and instrument monitoring	Ensuring data completeness and integrity for trusted use consumption.
Machine-readable data quality reports	
Instrument monitoring tools	
Data analysis and visualization	Analysis framework that includes visualization information techniques and automated data manipulations, such as data mining, feature tracking, reduction, etc. Server-side and in situ computation is necessary as the increase in data size and complexity of algorithms lead to data-intensive, compute-intensive challenges for CESD diagnostics, UQ, analysis, model metrics, and visualization.
UV-CDAT	
NCL, NCO, CDO, Matlab, IDL, VAPOR, R	
Open Source: R, EDEN (some versions), Py-ART	
Commercial: Matlab, IDL, etc.	
TECA (feature tracking)	
ACME diagnostics, PCMDI Metrics, iLAMB	
UQ: Dakota, PSUADE	
Collaboration and work management	To speed up, track, manage, and monitor key tasks, software, and infrastructure resources.
Confluence, JIRA, ServiceNow, Git, Pegasus	
Wiki	
NX technology to work on remote machines	
Citations and publications	Create unique data and user identifiers that link to data and metadata.
DOI tools (DOE OSTI) and ORCID Globus publication service	
Compute and storage facilities	HPC facilities deploy HPC systems, high-end storage, and data transfer nodes designed for accelerated scientific discovery.
ALCF, NERSC, OLCF	
Portal and search systems	Web-based user interfaces and content management systems for interactive tools and infrastructure use.
CoG, Drupal, WordPress D3, Solr, Elastic	
Workflow and provenance	Implemented APIs to capture workflow progress and provenance in infrastructure.
Swift, Tigres, Akuna, VisTrails, ProvEn, Jupyter	

Participants in the breakout session highlighted as key data capabilities the need for seamless unified search and access across data sets, UQ tools, and connection to a specific workflow. Server-side analysis and visualization, single sign-on/federated authentication, and tools to combine disparate data sets at different resolutions were also identified as important. This server-side analytics should consider the total costs of data-movement and analytics ease for users. A related need is flexible and scalable virtualized approaches that allow growth of the analytics over time. Virtualized and container-based approaches can enable new analytics functionalities to be added over time in a systematic manner. Further, they noted that provenance tools such as VisTrails need to be integrated with project-specific workflows.

Action items suggested by the group included building an inventory of tools in use by major projects, developing a strategy to integrate tools and service across facilities and infrastructures, providing tools as a service in the computing architecture, enabling a source code repository that is "common" with front-end-release via web browsers, and providing pre-created virtual machines (VMs) / Red-hat Package Mangers (RPMs) with a representative set of tools. Further, the group noted that structuring these needs and requirements in an actionable manner for computing and observational facilities is essential to success.

Contributors also detailed some potential methods for assessing tool maturity and their capabilities. Suggestions ranged from an App Store–style star rating/clearinghouse to publication references (DOI for tools, like Zenodo) to metrics tracking (number of contributors, most recent activity, number of diverse scientific projects the tool supporting, number of downloads/users/usage, etc.) to assessing the commitment level of developers to sustainability and software engineers to support.

Participants also discussed other action items related to existing tools and services benchmarking, such as software maintenance, security patches, connectivity to HPC resources, maintainability, installation, and documentation. One of these areas was the types of support expected by the science community. These expectations were diverse and included software documentation and user support, data quality issue and provenance communications support, community-used tools support, software maintenance, and deployment support.

Finally, the group tackled the topic of data and metadata conventions, such as climate forecast (CF) metadata conventions for model data and CF-type conventions for observations and experiments, and whether they should be adopted across many or all data centers. They agreed that common metadata, provenance, and DOI standards, including common assignment and collection approaches, should be developed, but that other community-followed standards such as HDF5, CSV, ISO should also be supported for broader data integration.

9. Data Services and Monitoring

The workshop included a discussion among participants about their data monitoring and computing and networking service needs within the proposed CESD integrated infrastructure. The session was particularly significant given the top-level requirement for increased reliability and resiliency of resources identified by the survey. Subsequent discussions of the survey results at the workshop supported the idea that scientists perceived the performance of existing resources as unreliable, in particular when used as part of more complex work processes across several resource types and/or institutions. However, exchanges in this broad group of workshop participants demonstrated that the domain scientists did not want to get involved in the operational aspects of the resources. Rather, they expect the facilities to provide easy-to-use, reliable services and identify and resolve issues proactively.

> Workshop participants requested a new class of monitoring services for the next generation of complex workflows. In particular, they want to see services that focus on capturing metrics on data and software downloads, users, and publications resulting from the reuse of their data and software by others.

For their part, the service providers identified a range of challenges that occur when supporting users in a distributed environment. Foremost is the challenge of exchanging comparable monitoring information across facilities. The use of software-as-a-service (SaaS) services such as Globus has been shown to improve overall system reliability by providing a reliable centralized location for problem detection, determination, and correction. PerfSONAR provides a network layer example of how such information sharing can help with early identification of potential problems and their solution or mitigation (transmit via a different route, store data in a different place in the interim, etc.), but it requires that service providers operate compatible monitoring services that capture similar information, as well as the ability to connect and evaluate the overall infrastructure health. Further, users and service providers identified the need for an event alert system that at different levels of detail informs infrastructure participants of issues. It was suggested that the working group investigate solutions developed by the LHC collaboration to manage their worldwide network of collaborating resources.

Infrastructure users, in turn, suggested a completely new type of monitoring services that they would like to see in a CESD integrated infrastructure. These services would be focused on capturing metrics on data downloads, data users, feedback on downloaded data, and publications resulting from the reuse of their data by others. In addition to data, they would like to see similar services for software tools that are shared throughout the infrastructure. The results of such metrics-capturing services should be available to both the data owners/software developers and data and software users. Discussions centered in particular on technologies and approaches that would support the tracking of data as it is analyzed and combined with other data products, to capture not just the bytes but also data re-use, impact, and attribution. Once

again, SaaS approaches have much to offer in this regard. The inclusion of DOIs as part of downloaded data products and automated insertion of acknowledgement sections were of particular interest.

Workshop participants requested a new class of monitoring services. Next to traditional service availability monitoring, in particular for complex workflows, they would like to see services that focus on capturing metrics on data and software downloads, users, feedback on downloaded data/software, and publications resulting from the reuse of their data/software by others.

10. Synergies with Peta- and Exa-scale Computing Hardware

In addition to local computing resources, climate and computational scientists are supported by high performance computing (HPC) facilities (i.e., ALCF, NERSC, OLCF) that deliver a balanced HPC environment with constantly evolving hardware resources and a wealth of HPC expertise in porting, running, and tuning real-world, large-scale applications. Currently, HPC facilities deliver multiple petaflops of compute power, massive shared parallel file systems, powerful data analysis and visualization platforms, and archival storage capable of storing many petabytes of data. This balanced hardware environment supports key collaborations between data infrastructure developers and HPC facility experts on the creation, debugging, production use, and performance monitoring of HPC parallel applications. A transition to exascale computing will bring energy efficient architectures with higher core counts and advanced data fabrics based on hierarchical memory technologies such as NVRAM. Data and flexibility-focused infrastructure, such as CADES and Argonne's Petrel and Magellan, when combined with HPC facility resources offer opportunities for leading edge techniques in data manipulation, storage and end user usability.

> LCFs have no policy for retaining data sets with a useful lifespan that extends beyond supported compute facility programs (e.g., the Computational Impact on Theory and Experiment [INCITE] program). In addition, lack of a single sign-on for authentication and federated access was also discussed as a hindrance in using multiple LCF computing hardware and resources.

CESD synergy with peta and exascale trends will hinge on leveraging technological advancement while maintaining a balanced computing environment that can support key collaborations between data infrastructure developers and HPC facility experts on the creation, debugging, production use, and performance monitoring of HPC parallel applications. The computing requirements of the CESD community are already tightly integrated into plans for future systems and continued dialogue can maintain those synergies.

Current major HPC facilities include petaflop systems featuring varied and disruptive HPC technologies, along with Lustre and GPFS file systems capable of storing petabytes of data. The computing infrastructure includes heterogeneous underlying hardware and software and cloud platforms to meet user needs, and employs large multi-core, multi-socket Linux clusters with a variety of processor types including GPUs. Partnering across the DOE SC and NNSA laboratories, the HPC facilities are preparing to launch several pre-exascale HPC systems set to bring hundreds of petaflops of computing power to the scientific community.

In the past decade, HPC facilities have deployed many Linux clusters containing thousands of nodes. Most clusters have similar commodity node-based architecture and provide a common programming model for ease of use. That is, they are built and maintained using commodity off-the-shelf hardware and open-source software. Node components are selected for performance, usability, manageability, and reliability. Most Linux clusters at DOE facilities run a common software environment based on Red Hat Linux with added kernel modifications, cluster system management, monitoring and failure detection, resource management, authentication and access control, development environment, and parallel file systems. Many of these components are developed and maintained in house; others are developed and

maintained in collaboration with a vendor partner. Looking ahead, more of the existing cluster-scale technologies will migrate into the compute node itself and renewed attention to the interconnects and memory hierarchies of which exascale systems will be comprised.

HPC and other DOE facilities deploy dedicated data transfer systems, called Data Transfer Nodes (DTNs) for moving data between facilities as required by science teams. In most cases, the DTNs are deployed in Science DMZ environments, which give the DTNs high-speed connectivity to the DTNs at other facilities and research institutions by means of the 100Gbps ESnet network, and run Globus software for convenient access to high-speed data movement capabilities. Collaboration between the HPC facilities and ESnet (and other network organizations) are working to instrument the hardware with application software (such as with ESGF and Globus) to develop large-scale and reliable disk-to-disk data transfers. This software will allow for isolated sandboxes and workflow substrates for experiments and different scientific workflows.

The hardware system effort also combines traditional HPC with emerging cloud technologies. More specifically, these platforms use (1) virtualized high-speed Infiniband networks, (2) a combination of high-performance file systems and object storage, (3) diverse analytics infrastructures including graph engines and memory intensive computing platforms, and (4) virtual system environments tailored for data intensive, science applications. There is growing attention to configuring these analytics environments to be cognizant of the data-analytics application needs. For example, systems are increasingly set up so that the memory and storage hierarchy provides the capability to perform data-proximal processing. Surrounding the data storage is a cloud of HPC resources with many processing cores and large memory coupled to the storage through high-speed network backplanes. Virtual systems can be tailored to a specific scientist and provisioned on the compute resources with extremely high-speed network connectivity to the storage and to other virtual systems (see **Figure 7**).

Finally, in addition to large-scale data analysis, systems are being used for hosting large-scale data services, such as ESGF node services at multiple locations around the globe. The data that are stored within the federated infrastructure includes simulation, observation, and reanalysis data for multiple inter-comparison projects. We give examples for a limited-scale deployment with a constrained scope. These resources would be considered as compute and storage building blocks for larger analytics needs and can be scoped to scale out according to the needs of the program.

Table 5. *Examples of HPC facility component hardware system capabilities, description, and configuration. Note these are scoped as building blocks that can be scaled according to needs.*

Capability and Description	Sample Analytics Configuration
Persistent Data Services Virtual machines or containers deployed for web services. Examples include ESGF, GDS, THREDDS, and FTP.	8 nodes with 128 GB of RAM, 10 GbE, and FDR IB
Database High available database nodes with solid-state disk.	2 nodes with 128 GB of RAM, 3.2 TB of SSD, 10 GbE, and FDR IB
Remote Visualization Enables server-side graphical processing and rendering of data.	4 nodes with 128 GB of RAM, 10 GbE, FDR IB, and GPUs
High Performance Compute Several 1,000 cores coupled via high-speed Infiniband networks for elastic or itinerant computing requirements.	~100 nodes with 32 to 64 GB of RAM, and FDR IB
High-Speed/High-Capacity Storage Petabytes of storage accessible to all the above capabilities over the high-speed Infiniband network.	Several storage nodes configured to support PBs of RAW spinning disk and object store capacity
Long-term/Persistent Tape Storage Tens of petabytes of long-term storage that is accessible upon request. Data is staged to disk cache and use is notified when requested data are retrieved.	50 PB (or more) of high-performance storage system tape archive

Geographically Distributed High-Speed/High-Capacity Storage Many petabytes of high reliability storage distributed across physical locations allowing for irreplaceable and high value data to be stored more cost effectively.	10 PB (or more) of high-reliability storage per site across several sites

11. Network Services

High-speed network services will enable fast and robust connections to be made between participating DOE laboratories (including HPC facilities), NASA, NOAA, NSF, and international federated data centers, effectively transporting hundreds of petabytes of large-scale simulation and observation data. As an example, collaborating centers utilize GridFTP for data replication and backup, driven by Globus. These network services also use the national and international 100Gbps Internet connections provided by ESnet, Internet2, and other domain-specific networks. The International Climate Network Working Group (ICNWG), a working group of the ESGF, is engaged in an effort to improve and sustain the data transfer performance between major climate data centers in support of data replication and data transfer efforts (**Figure 8)** [ICNWG 2015]. The ultimate goal of this effort is to achieve managed sustained disk-to-disk throughput of multi-petabyte data sets between the centers for replication. Achieving this capability will in addition allow the infrastructure to meet the heavy demands of moving large-scale data to centers for critically important compute operations such as federated uncertainty quantification calculations and ensembles.

Current high-speed reliable data movement is not sufficient for CESD data resiliency and backup needs.

Figure 8. *International Climate Network Working Group (ICNWG).*

With the advent of software-defined networking, a rich set of APIs for interacting with the network, such as setup and route direction is possible. The data grid can program the switches to use disjoint routes when doing multi-stream large-data transfers for replication and/or federated computing.

For network performance measuring, perfSONAR could be integrated into the infrastructure. perfSONAR measures the network performance capabilities at the end sites by using the tools bandwidth test controller (for throughput testing, run every few hours) and One-Way Active Management Protocol (low-bandwidth one-way delay measurement and packet loss testing, running continuously). The results could be stored on a server, which can be viewed using an API or Web browser.

To monitor network performance and services, a perfSONAR node (a virtual machine) must be deployed alongside participating standard nodes as representatives of that host environment. To maximize network services, a number of perfSONAR boxes will be installed within the infrastructure spanning the federated data centers and network domains. This will immediately help to address and troubleshoot local-area and wide-area network issues.

12.Participation with Broad Multi-Agency Data Initiatives

> The CESD workshop attendees understand that in order to be successful they must strengthen their partnership with other national and international agencies.

As DOE considers the design and implementation of a broad capability for data and informatics in service of its climate and environmental science missions, it is imperative to catalog existing and emerging capabilities across multiple institutions and agencies, including international efforts, and determine how best to integrate new and existing capabilities. The development of a robust predictive understanding of Earth's climate and environmental systems is an inherently interdisciplinary problem. An integration of observational and experimental data, process knowledge, and predictive modeling across a wide range of traditional science domains, including physical, biological, and sociological, is necessary for development of sustainable solutions to pressing energy and environmental challenges. As DOE pushes forward to fully engage with these challenges, a broad perspective on current and emerging data and informatics systems and their capabilities will provide the best opportunity for deep collaboration and rapid progress toward a system that serves agency needs while improving Earth science understanding for the whole community.

Major Earth science data and informatics systems and services are already operational in other U.S. agencies, including large-scale efforts at the National Aeronautics and Space Administration (NASA), National Oceanic and Atmospheric Administration (NOAA), and National Science Foundation (NSF). Those efforts are summarized below as an illustration of the depth and scope of current and emerging efforts in this area. A critical first step in advancing DOE's capability in this area should be to conduct a much more intensive technical review of these and other existing programs. A wide variety of tools and technologies are in use, many of which are well evolved and could serve a beneficial role in a DOE system. Significant capabilities developed with DOE support are also available, as described above.

EOSDIS provides end-to-end capabilities for NASA Earth science data from multiple sources, including satellites, aircraft, and field measurements (https://earthdata.nasa.gov). The Earth Science Data and Information System (ESDIS) project manages the science systems of the EOSDIS, providing science data to a wide community of users for NASA's Science Mission Directorate. Major ESDIS capabilities and objectives include: processing, archiving, and distributing satellite data; providing tools for archiving, processing, and distributing of a variety of Earth science data; ensuring ready access to data promoting research in the areas of climate and environmental change, guided in part by the gathering and analysis of data user metrics; and promoting interdisciplinary data use.

ESDIS supports twelve Distributed Active Archive Centers (DAACs, https://earthdata.nasa.gov/about/daacs), as well as many Science Investigator-led Processing Systems (SIPSs). A general view of ESDIS data flow is from primary instrumentation to a dedicated SIPS, where raw instrument data are processed to produce Earth Observing System (EOS) standard products, and from the SIPS to the relevant DAAC for distributing, archiving, and performing a broad range of user services, including some value-added product generation and web-based access and analysis tools. Given the diversity of raw data sources (satellite, aircraft, field measurement) and science domains of interest among the various SIPSs and DAACs, a coordinated strategy for documented interfaces has been an essential element in smooth operation of ESDIS. Design documents and interface control documents with

standardized formats and information content are in place at each step in the data lifecycle, from instrument to DAAC and out to science users.

The EOSDIS data strategy includes a unified approach to gathering, indexing, and accessing metadata across all of its products, investigator-led teams, and data centers. The emerging metadata framework within EOSDIS is called the Common Metadata Repository (CMR), which brings together previously developed capabilities from the Global Change Master Directory (GCMD) and the EOS Clearing House (ECHO). CMR will validate metadata adhering to various standards (e.g. ISO19115, GCMD DIF) against a common standard (the Unified Metadata Model).

With numerous data centers (DAACs) and upstream service providers (including SIPSs), the integrated ability to search across the entire EOSDIS holdings is a crucial performance metric for ESDIS. The Earthdata Search application provides flexible keyword searching as well as a range of data discovery tools and services. Tools include web clients for browsing and ordering of mapped data sets, including time varying data and open-source geospatial analysis tools (e.g. region-of-interest subsetting, reprojection, and geolocation). Another recently developed tool is the Global Imagery Browse Service, which helps to solve the problem of many data sets being delivered in small "granules" that must be stitched together in space and/or in time before arriving at first-look evaluations. ESDIS also provides a system for serving large and complex data sets to a broad range of users for near-real-time applications (the Land, Atmosphere, Near real-time Capability for EOS, or LANCE).

The collection of discipline-oriented DAACs is designed and operated as a distributed data and informatics system, with coordination managed through well-defined interfaces and standards. A special ESDIS Standards Office (ESO) provides coordination for the list of standards approved for use in NASA Earth Science Data Systems and community organization through teleconferences and working groups for discussion of existing and emerging standards. The DAACs as a whole provide and/or deploy a wide array of data discovery tools. In addition to tool sets and capabilities already mentioned, there are numerous data visualization and analysis tools supporting a wide variety of data types and sources. Examples include Giovanni (Giovanni.gsfc.nasa.gov) and MODIS subsetting (http://daac.ornl.gov/MODIS) and overlay tools, with an emphasis on multivariate and multi-temporal remote-sensing data products. EOSDIS currently includes over 8,000 unique data collections, with a total archive volume of 9 PB, growing at over 6 TB/day. The system registers over 2 million distinct users, with an average end-user distribution volume of 28 TB/day (statistics as of Sept. 2014, https://earthdata.nasa.gov/about/system-performance).

EOSDIS participates in a number of national and international data community collaborations, including the Federation of Earth Science Information Partners (ESIP Federation), U.S. Group on Earth Observations (USGEO), and Open Geospatial Consortium (OGC). EOSDIS actively participates in and supports the U.S. government's Climate Data Initiative (www.data.gov/climate) and Big Earth Data Initiative.

NOAA provides an integrated view of climate and weather data at multiple scales from regional to global through their Climate.gov project (www.climate.gov), which began in 2010 as a prototyping collaboration among four NOAA offices (Climate Program Office, National Climatic Data Center, Coastal Services Center, and Climate Prediction Center). The "Maps and Data" section of Climate.gov is developing to support storage, retrieval, and graphical presentation of climate and weather-related data from across NOAA and its partners' data centers. Science and data panels guide the evolution of Climate.gov, with membership from within NOAA and from universities and other agencies. The data panel, which includes senior data managers from major Earth system data centers, provides input on available data sets and current and emerging technologies for data search and delivery. A relatively small number of well-curated data sets are presented with great attention to graphical formats and clear documentation, targeting a broad audience of science, policy, and education users.

NOAA has recently merged three major data centers (National Climatic Data Center, National Geophysical Data Center, and National Oceanographic Data Center) into a single distributed system, the National Centers for Environmental Information (NCEI). Atmospheric, oceanographic, coastal, and geophysical data products and services are being organized using a common set of data service technologies and provided through a common set of interfaces. Coverage includes data products at both national and global scales, and NCEI services target a broad user base in research and application areas. NCEI partners with Climate.gov, NOAA's National Weather Service (Weather.gov), the National Integrated Drought Information System (NIDIS, www.drought.gov), and the U.S. Global Change Research Program (USGCRP, www.globalchange.gov).

Other parts of NOAA support additional data services, such as the National Centers for Environmental Prediction (NCEP), which maintains and distributes a wide range of climate-relevant information, and the Geophysical Fluid Dynamics Laboratory (GFDL), which supports search, retrieval, and distribution of climate modeling data, including implementation of an ESGF node.

In response to the Big Data Initiative announced by the White House Office of Science and Technology Policy in 2012, NSF has invested in multiple efforts, including the Data Observation Network for Earth (DataONE), EarthCube, and a project integrating Algorithms, Machines, and People (AMP).

DataOne (www.dataone.org) is intended to provide a single point of access to a broad range of data resources, drawing together a meta-collection of Earth data from a large number of partners. A working group structure is used to provide guidance on current and emerging efforts connected to the lifecycle of large and complex data systems. Working groups currently include Sustainability and Governance, Community Engagement and Outreach, Cyberinfrastructure, and Usability and Assessment. Data search capabilities link users to one or many of the 27 (currently) member "nodes." In addition to data access, DataONE also provides and updates detailed information on best practices for data management and maintains a compilation of useful software tools. DataONE partners with the Data Management Planning Tool (DMPTool, dmptool.org) to provide resources for creating, reviewing, and sharing data management plans.

EarthCube, supported by both Geosciences and Advanced Cyberinfrastructure programs in NSF, seeks to increase the availability of data and associated tools and services in the broad Earth sciences community, increasing knowledge availability for society as a long-term goal.

AMP (amplab.cs.berkeley.edu) addresses scientific challenges related to the application of newly available large-scale computing resources to the burgeoning volume of data and growing requirements for data analysis. Variable data quality, formats, and sources make it difficult to apply traditional analysis algorithms to the largest data sets, and the available computer architectures are not always compatible with current algorithmic and programming models. Machine learning, data mining, language processing and speech recognition are all areas being explored under AMP as avenues for improved knowledge discovery.

The large data and informatics efforts summarized above are just a few of many efforts currently underway in this domain. A comprehensive list is beyond the scope of the workshop or this reporting but would include dozens of agencies and institutions at the local, state, national, and international levels. Beyond developing a more complete awareness of this broad landscape and a refined appreciation for the capabilities and expertise available in different agencies and centers, it is also necessary to define strategic partnerships that meet DOE BER objectives while providing an added value to a broad and growing data and informatics community. Some of this coordination will take place at the level of agency and organizational representatives, but there is also a role for data management practitioners and data center operations specialists, in coordination with science team representatives across a range of projects and agencies, to develop system requirements and suggest creative adaptations and/or reconfigurations of existing efforts to meet those requirements. If these integration efforts can reach across agency and

institutional boundaries, it seems likely that efficiencies of scale and leveraging of unique capabilities will emerge. The present workshop report should be seen as one step toward the realization of that broader objective.

13.References

[ACME 2014]
ACME Council, 2014 Accelerated Climate Modeling for Energy: Project Strategy and Initial Implementation Plan, http://climatemodeling.science.energy.gov/sites/default/files/publications/acme-project-strategy-plan.pdf.

[AmeriFlux 2015]
AmeriFlux Network Management Center home page. http://ameriflux.ornl.gov/.

[ARM 2015]
ARM Climate Research Facility home page. http://www.arm.gov/.

[ASCAC Data Report 2013] DOE ASCAC Data Subcommittee Report, "Synergistic Challenges in Data-Intensive Science and Exascale Computing," technical Report, U.S. Department of Energy Office of Science, March 2013. http://science.energy.gov/~/media/40749FD92B58438594256267425C4AD1.ashx.

[BER VL 2015]
Janet Braam, Judith A. Curry, et al., BER Virtual Laboratory: Innovative Framework for Biological and Environmental Grand Challenges. Office of Biological and Environmental Research, Office of Science, Department of Energy, http://genomicscience.energy.gov/program/beracvirtuallab.shtml.

[CDIAC 2015]
Carbon Dioxide Information Analysis Center home page. http://cdiac.esd.ornl.gov/.

[Chard et al. 2014]
Chard, K., Tuecke, S. and Foster, I. Efficient and Secure Transfer, Synchronization, and Sharing of Big Data. *Cloud Computing, IEEE*, 1(3):46-55, 2014.

[Chard et al. 2015]
Chard, K., Pruyne, J., Blaiszik, B., Ananthakrishnan, R., Tuecke, S. and Foster, I., Globus Data Publication as a Service: Lowering Barriers to Reproducible Science. 11th IEEE International Conference on eScience Munich, Germany, 2015.

[Cinquini 2014]
Luca Cinquini, Daniel J. Crichton, Chris Mattmann, John Harney, Galen M. Shipman, Feiyi Wang, Rachana Ananthakrishnan, Neill Miller, Sebastian Denvil, Mark Morgan, Zed Pobre, Gavin M. Bell, Charles M. Doutriaux, Robert S. Drach, Dean N. Williams, Philip Kershaw, Stephen Pascoe, Estanislao Gonzalez, Sandro Fiore, Roland Schweitzer, "The Earth System Grid Federation: An open infrastructure for access to distributed geospatial data," Future Generation Computer System 36: 400–417, Volume 36, July 2014, DOI: 10.1016/j.future.2013.07.002.

[ESGF 2015]
The Earth System Grid Federation home page. http://esgf.llnl.gov/.

[ESS Report 2015] U.S. DOE, 2015, Environmental System Science (ESS) Workshop on Model-Data Integration: Modeling Frameworks, Data Management and Scientific Workflows, DOE/ SC-0178. U.S. Department of Energy Office of Science. http://doesbr.org/ESS-WorkingGroups.pdf.

[ICNWG 2015]
The International Climate Network Working Group home page (2015): http:/icnwg.llnl.gov/.

[ScienceDMZ 2015]
Eli Dart, Lauren Rotman, Brian Tierney, Mary Hester, and Jason Zurawski, "The Science DMZ: A Network Design Pattern for Data-Intensive Science", SC13: The International Conference for High Performance Computing, Networking, Storage and Analysis. Denver CO, USA, ACM. DOI:10.1145/2503210.2503245, November 19, 2013, LBNL 6366E. http://fasterdata.es.net/science-dmz/.

[Williams 2014]
Dean Williams, Giri Palanisamy, Galen Shipman, Thomas Boden, and Jimmy Voyles, "Department of Energy Strategic Roadmap for Earth System Science Data Integration," Big Data, 2014 IEEE International Conference Proceedings: pages 772–777, Washington D.C., 27–30 October 2014, DOI:10.1109/BigData.2014.7004304.

Appendix 1: Attendees Findings

The workshop participants have reviewed current practices and future plans for multiple CESD science projects in the context of the challenges facing both the Virtual Laboratory and data infrastructure. The findings drew from workshop presentations, workshop reports, expert testimony, and use cases. Data-intensive activities are increasing in all CESD science endeavors, and HPC compute facilities are a key enabler of these activities. We briefly summarize below key findings from the attendees from the perspective of identifying investments that are most likely to positively impact both CESD's science goals and mission space.

Topic	Finding
Interagency Partnership	The challenges of distributed big data management and analysis are too large for CESD to solve alone. CESD will only succeed in its exciting Virtual Laboratory goals by leveraging best-of-breed research data management technologies used by the science community.
Reproducibility and Repeatability	The sheer size of current and expected future archives makes it impossible to store and analyze data on the users' personal workstations. Therefore, there is the need to submit complex data analysis workflows that seamlessly process data that are stored at distributed locations. The detailed workflow metadata (inputs, outputs, algorithms) must be provenanced captured and made publicly available so that other researchers can fully understand and reproduce/repeat the results.
Funding of Resources	DOE researchers have led the world in the application of advanced computing to computational simulation. In contrast, DOE climate and environmental science suffers from an ad hoc, under-resourced research data infrastructure. This situation significantly hinders progress in research programs of great scientific and societal importance. For example, availability and reliability of hardware and other resources for data analysis is a major issue.
Storage and computing	CESD currently lacks the storage and computing resources required to achieve its science goals. CESD should establish strong strategic partnerships with ASCR to ensure availability of those resources, and examine the feasibility of using commercial cloud resources for some purposes.
Scalability	Services must be designed to be able to scale to the order of magnitude of future data and metadata archives that are expected in the next 5–10 years, while still guaranteeing a satisfactory level of performance to the users. In particular, the infrastructure must be able to support the hundreds of petabyte-sized distributed archive that is expected to be generated by the next generation of climate models and higher resolution observing instruments.
Proactive Engagement with CESD Projects	Projects and development teams must continuously and proactively engage with all possible areas of the project/programs—data users, data providers, project coordinators, infrastructure providers, and funding program managers. This will guarantee that the data, software, and resouces are developed and utilized to fulfill the stakeholders' requirements, maximizes the users' satisfaction, and achieves the expected level of service.
Model Runs	Model development and modelers have varied data management needs, which includes performing many small model runs with rapid turnaround during the model development phase, more computationally demanding uncertainty quantification and optimization work for model refinement, and massive data runs on leading supercomputers with the full array of analysis, diagnostics, and model metrics features once the models are in production. Modelers are expect to utilize shareable reproducible/repeatable workflows, access data from many different heterogeneous data sources, and run HPC in situ analysis, diagnostics, and model metrics.
Data Transfers	When conducting large-scale analysis of data sets from multiple climate models, the data sets are typically assembled at the HPC facility where the scientist has the necessary computing allocation to run the analysis. This requires high-performance data transfer capabilities between the major data centers and the HPC facilities, which have the necessary computing and storage capabilities to support these large-scale in situ analyses. It is therefore critical that the data centers and HPC facilities support the transfer of large-scale data sets to major computing facilities in addition to data subsetting and co-located data analysis services. At present, researchers spend an enormous amount of time thinking about where data is physically and how to co-locate data physically for analysis.
Uncertainty Quantification	A system that cross-references uncertainty estimates on observational and modeling results is

(UQ):	needed to ensure that empirical constraints are applied appropriately. Analysis of large and multi-dimensional model outputs is required to interpret UQ results. Filtering of sensitivity analysis results produces a reduced set of parameters for formal estimation, but these results can vary in space and time, placing high demands on the analysis framework and requiring engagement of expert knowledge.
Data access and Ownership	Scientific projects, data providers and users expect a reliable data infrastructure to make their products visible and accessible, while being able to control and track utilization and receive appropriate credit for their contributions. Data should be clearly identifiable and recognizable via DOI's and owners of the data must be recognized, possibly by ORCIDs.
Discovery	It is difficult to find physically related data between programs and projects in CESD scientists.
Standardization	Whenever possible, the infrastructure must conform to established standards for flexible interactions. This will also maximize interoperability with other agencies' (i.e., NASA, NOAA, NSF, international) data systems and facilities. Additionally, interoperability greatly increases the level of user satisfaction, as users are not compelled to learn and develop different techniques to access services from different systems.
Data Movement	High-speed, reliable data movement is essential to Virtual Laboratory goals. CESD should work closely with ESnet and Globus to ensure high-speed, reliable, and secure end-to-end communications between its researchers, its facilities, and other relevant resources.
Best Data and Software Practices	Software and service reliability is frequently underemphasized in science, but is absolutely essential to Virtual Laboratory goals. CESD must ensure that future programs leverage best-practice methods to achieve the high reliability required to meet science goals. This means developers must strive to apply recommended best practices in all phases of the data and software lifecycle (design, development, testing, deployment, operation) and across all software layers (**Figure 2**). This can be achieved by many collaborative events such as software code sprints, code reviews, and test coverage analysis. This also includes common data curation policies across CESD and inter-agencies.
Search and Discovery	A user must be able to search, discover, download, and analyze data hosted at different centers and facilities as if they were served from a single location. The distributed nature of the system must be totally transparent to end users and clients. This means establishing common metadata for raw and post-processed data across CESD to facilitate search and discovery.
Monitoring and Metrics	To permit continuous, data-driven improvement of Virtual Laboratory operations and investments, the VL should incorporate extensive monitoring and logging capabilities to permit detailed and accurate analysis of VL performance, reliability, security, and usage. This also include facilities for capturing and analyzing metrics about utilization of services, as well as for estimating the impact of the data infrastructure over the science community (for example, as quantified by the number of science papers that use some data sets downloaded, or based on processing algorithms executed on LCF servers). These metrics can be used to both improve the performance and quality of services and for reporting usage to the CESD program mangers.
Modularity	The infrastructure must not be built as a monolithic package that must be installed and upgraded as a whole. Rather, it should be based on the integration of several servers and libraries that are meant to be upgraded and possibly replaced individually (**Figure 2**). This philosophy enables the infrastructure to continuously evolve to incorporate new advances in all classes of services: data discovery, transfer, analysis, visualization, etc.
Local or Remote and In situ Analysis	Server-side and in situ computation is necessary as the increase in data size and complexity of algorithms lead to data-intensive, compute-intensive challenges for diagnostics, UQ, analysis, model metrics, and visualization. For complete flexibility, the analysis system must abstract away the data file's phyiscal location and let the back-end dynmamic resource manager decide how, when, and where to move the data for small- and large scale analysis. This includes the creation of a cloud-based CESD anlysis platform which can scale to the needs of CESD scientists.
Unified Access Control	In particular, a user or client must not be asked to authenticate or be authorized separately at all data centers or HPC facilities. Rather, the system infrastructure must support Single Sign-On for authentication and federated access control, whereby the authorization statements issued by one center are honored by the other peer centers, for accessing the same class of resources.

Appendix 2: Workshop Example Questions

Data Infrastructure	
	How do we integrate all CESD and eventually all BER data holdings?
	What are the missing components that need to be developed to integrate existing BER data archives?
	What type of construct do we use (e.g., co-located, federated)?
	Should this integrated environment construct be a facility or a project?
	Can this construct be complimentary to existing data efforts supported by other agencies (e.g., EarthCube, NASA-DAAC, etc.), and if so how?
	How do we serve the data to our communities? • What modes of data transfer should be available to the users of this system?
	Should a simple compute visualization framework be incorporated? • What should its capabilities be? • Are these calculations done locally or server-side
Compute Environment	
	How will BER scientists be doing research and interacting with the large volumes of data 10+ years from now?
	What type of data and computing environment will be necessary for this seamless integration?
	Will it be possible within a heterogeneous compute environment to support this type of system?
	Will task automation be a necessary component of this system?
	How will code reusability be addressed within this construct?
	Will exascale compute resources be a necessary component or a complimentary resource for this structure? • Regardless of where this system is implemented, it must appear transparent to a user. What necessary components must be addressed to make this happen? • Which components are key failure points?

Appendix 3: Survey Questions – Overall Ranking

The new DOE mandate on data management and sharing has clearly penetrated the community and raises questions for many. This is borne out by high scores for a number of related questions such as:

- Easy way to publish and archive data using one of the DOE data centers. This question received the highest score overall, with 4.79 average rating. Nearly 70% of responders identified this as highest or second highest requirement.
- User support for data access and usage achieved a high rating of 4.64.
- Access to enough computational and storage resources raised similarly high interest in this context, with 4.52/41%.

One could also tie the increased interest in collaborative environments for the sharing of data and information within and between scientific groups into this topic area.

Survey Question	Average Rating
Easy way to publish and archive your data using one of the DOE data centers	4.79
Means for comparison of diverse data types generated from observation and simulation	4.71
User Support for data access and usage	4.64
Access to sufficient observational and experimental resources	4.58
Access to enough computational and storage resources	4.52
Ingest and access to large volumes of scientific data (i.e., from data archive to super computer)	4.49
Quality control algorithms for data	4.46
Would you like a unified and single user account to access all BER and ASCR resources?	4.44
In-situ analysis of observational, experimental and computational results: the ability to interpret results and verify new insights within the context of existing scientific knowledge	4.4
Means of comparison of data collected at different scale	4.34
Collaborative environments	4.31
Availability of ancillary data products such as data plots, statistical summaries, data quality information, and other documentation	4.22
Rapid data quality assessment during discovery	4.18
Interoperability: Interfaces that ensure a high degree of interoperability at format and semantic level between repositories and applications	4.18
Data manipulation before download (averaging, subsetting, etc.)	4.16
Provenance capture information for data	4.11
Reproducibility	4.06
Across institutions and communities: Libraries, repositories that allow for community-wide authentication and access	4.06
Improved user interfaces	4.0
Unified data discovery for all BER data sources to support your research	4.0
Software that enables small teams to engage in large scale ensemble and UQ simulations	3.88

Survey Question	Average Rating
Direct data delivery into ASCR computing systems from BER data resources	3.86
Software to ensure workflow resilience and recovery from errors	3.85
Data visualization tools	3.85
Real time data quality control during data collection	3.74
Support for the creation of scientific workflows	3.68
Data intention: Methods and languages for describing and adhering to intellectual property in systems where not all the data is openly available	3.57
Real-time access to life data streams	3.25
New techniques to work with deep memory hierarchies on extreme scale computing systems	3.24

Appendix 4: Workshop Agenda

Time	Topic
Thursday August 13, 2015	
8.45 am - 9.10 am	**Welcome and introduction** (Gary Geernaert) **Workshop charge** (Jay Hnilo)
9.10 am– 9.30 am	**Survey responses** (Kerstin Kleese-Van Dam)
9.30 am – 9.45 am	**Identifying CESD computational and data environment** (Dean Williams, Giri Palanisamy)
9.45 am - 10.00 am	**Break**
10.00 am– 11.00 am	**Science Drivers** Discussion Lead (Peter Thornton) ● Example use case requirements (Jay Hnilo) 10 mins ● Define what are the key things that are difficult to do today and are impeding scientific progress or productivity ● Science case discussion (50 mins) (list science drivers; HW assignment convert science drivers to use cases)
Breakout sessions	
11.00 am– 12.30 pm	**Data Services to Support Science Requirements** Red Team: Discussion Lead (Forrest Hoffman) Blue Team: Discussion Lead (Shaocheng Xie) Questions: ● What are the key challenges that scientists encounter? ● What data services would address the identified challenges? What exists already today? What do we still need? What are the key characteristics that these services need to have to be successful (i.e. integrated, easy to customize etc.)? ● What are the key impediments (on the data provider/service provider side) in delivering these services? ● Which services should be developed with the highest priority and what would be their measurable impact on science?
12.30 pm - 1.30 pm	**Lunch**
1.30 pm- 2.30 pm	Breakout Session Reports and discussion: 30 minutes per team
2.30 pm- 4.00 pm	**Required Data Center and Interoperable Services** Red Team: Discussion Lead (Margaret Torn) Blue Team: Discussion Lead (Tom Boden) Discuss top priority services required to meet the communities need as part of an integrated infrastructure, including topics such as: ● Data integration and advanced metadata capabilities ● Data and metadata collection and sharing capabilities ● Data quality, uncertainty quantification, and ancillary information ● Use of broader ontology for discovery and use of CESD data sets ● Data discovery and access, data downloading and subsetting services and capabilities ● Data preparation services and tools ● Authentication and security ● Local and remote publication services ● Local and remote catalog and search services, data transfer services ● Human computer interface (i.e., User Interface, APIs, etc.) ● Resource discovery and allocation services ● Workflow services (link together scientific or project execution)

	• Computing services • Exploration services (includes analytics and visualization) • Identify key gaps, identify benefitting communities, and prioritize
4:00 pm- 4:15 pm	**Break**
4:15 pm- 5:15 pm	Breakout Session Reports and discussion: 30 minutes per team
Friday, August 14, 2015	
8:30 am- 10:00 am	**Advanced Computational Environments and Data Analytics** Red Team: Discussion Lead (Scott Collis) Blue Team: Discussion Lead (Paul Durack) Questions: • What are the key challenges that scientists encounter? • What capabilities would address the identified challenges? What exists already today? What do we still need? • What are the impediments for resource providers and software developers to provide these missing capabilities? • Which requirements need to be addressed with the highest priority and what would be their measurable impact on science? Possible discussion topics: • Define a scalable compute resource (clusters and HPCs) for CESD data analysis • Data analytical and visualization capabilities and services • Analysis services when multiple data sets are not co-located • Performance of model execution • Advanced networks as easy-to-use community resources • Provenance and workflow • Automation of steps for the computational work environment • Resource management, Installation and customer support • Identify key gaps, identify benefitting communities, and prioritize
10:00 am - 10:15 am	**Break**
10:15 am - 10:45 pm	Breakout Session Reports and discussion: 15 minutes per team
10:45 am - 11:45 pm	**Inventory of existing CESD data tools and services, benchmark of tools for potential reuse** Red Team: Discussion Lead (Deb Agarwal) Blue Team: Discussion Lead (Jennifer Comstock) Suggested subtopics: • What tools have been identified during the previous discussions that should be made more widely accessible to the CESD community? • What other tools are there that could address key needs? • How should tools and services be made available today and in the future in an integrated infrastructure? What level of support would be expected from the science community? • How do we want to assess the maturity and capability of tools (e.g. benchmarks or crowdsourcing)? • Are there any conventions that are needed for your project?
11:45 am - 12:15 pm	Breakout Session Reports and discussion: 15 minutes per team
12:15 pm - 1:15 pm	**Lunch**
1:15 pm- 2:00 pm	General discussion: Data Services and Monitoring Discussion Lead (Eli Dart) Questions: • What level of service, monitoring, maintenance and metrics needed for data services and tools? • What do service providers want to see from others? • What do the scientists want to have access too?

2:00 pm - 2:30 pm	**General discussion: Participation with broad/multi-agency data initiatives** Discussion Lead (Peter Thornton) Suggested subtopics: ● Standards and services that needs to be adopted within the compute environment that will allow CESD to participate in multi-agency data initiatives such as EarthCube, USGEO etc. ● Data sharing with NASA DAACs, NOAA, and other agencies
2:30 pm- 3:00 pm	**Summary of action items, workshop report draft** **Follow up and future workshop ideas**

Red Team Members:
David Bader (LLNL) (Modeler)
Forrest Hoffman (ORNL) (Modeler)
Deb Agarwal (LBNL) (Data Management)
Rob Jacob (ANL) (Data Scientist)
Timothy Scheibe (PNNL) (Data Management)
Margaret Torn (LBNL) (Data Scientist)
Andy Vogelmann (BNL) (Modeler)
David Skinner (LBNL) (Data Center)
Scott Collis (ANL) (Data Scientist)

Blue Team Members:
Phil Rasch (PNNL) (Modeler)
Paul Durack (LLNL) (Data Scientist)
Peter Thornton (ORNL) (Modeler)
Michael Wehner (LBNL) (Data Scientist)
Tom Boden (ORNL) (Data Management)
Jennifer Comstock (PNNL) (Data Scientist)
Shaocheng Xie (LLNL) (Data Scientist)
Mallikarjun Shankar (ORNL) (Data Center)
Eli Dart (ESnet) (Data Center)

Appendix 5: Workshop Participants

Name	Area of Representation	Affiliation	E-mail Address
Participants			
Argawal, Deb	ESS	LBNL	daagarwal@lbl.gov
Bader, David C.	ACME, LLNL Climate Science	LLNL	bader2@llnl.gov
Boden, Thomas A.	Ameriflux, CDIAC, FACE, NGEE	ORNL	bodenta@ornl.gov
Collis, Scott M.	HPC, Py-ART, Radar	ANL	scollis@anl.gov
Comstock, Jennifer	ARM, ASR	PNNL	jennifer.comstock@pnnl.gov
Dart, Eli	ESnet	ESnet	dart@es.net
Durack, Paul J.	PCMDI, MIPs, RGCM	LLNL	durack1@llnl.gov
Hoffman, Forrest M.	ILAMB, ACME	ORNL	hoffmanfm@ornl.gov
Jacob, Robert	HPC, ACME	ANL	jacob@mcs.anl.gov
Kleese van Dam, Kerstin *	EMSL, ARM	PNNL/BNL	kerstin.kleesevandam@pnnl.gov
Palanisamy, Giriprakash *	ARM, NGEE	ORNL	palanisamyg@ornl.gov
Rasch, Philip J.	ACME	PNNL	philip.rasch@pnnl.gov
Scheibe, Timothy	EMSL	PNNL	tim.scheibe@pnnl.gov
Shankar, Mallikarjun	OLCF	ORNL	shankarm@ornl.gov
Skinner, David	NERSC	LBNL	deskinner@lbl.gov
Thornton, Peter	ACME, NGEE	ORNL	thorntonpe@ornl.gov
Torn, Margaret S.	Ameriflux, ASR	LBNL	mstorn@lbl.gov
Vogelmann, Andrew		BNL	vogelmann@bnl.gov
Wehner, Michael F.	CASCAD	LBNL	mfwehner@lbl.gov
Williams, Dean N. *	ACME, MIPs, ESGF	LLNL	williams13@llnl.gov
Xie, Shaocheng	ACME, ARM, RGCM/ASR (CAPT)	LLNL	xie2@llnl.gov
Participants from DOE Program Offices			
Bayer, Paul	DOE BER Program Manager	BER	paul.bayer@science.doe.gov
Geernaert, Gary	DOE BER CESD Director	BER	gary.geernaert@science.doe.gov
Hnilo, Justin *	DOE BER Program Manager	BER	justin.hnilo@science.doe.gov
Joseph, Renu	DOE BER Program Manager	BER	renu.joseph@science.doe.gov
McFarlane, Sally	DOE BER Program Manager	BER	sally.mcfarlane@science.doe.gov
Ndousse-Fetter, Thomas	DOE ASCR Program Manager	ASCR	thomas.ndousse-fetter@science.doe.gov
Petty, Rickey	DOE BER Program Manager	BER	rick.petty@science.doe.gov

* Workshop and report co-chairs and organizers

Appendix 6: Acronyms

Acronym	Description
ACME	Accelerated Climate Modeling for Energy: DOE's effort to build an Earth system modeling capability tailored to meet the climate change research strategic objectives
ALCF	Argonne Leadership Computing Facility, sponsored by DOE (http://www.alcf.anl.gov)
AmeriFlux	AmeriFlux Site and Data Exploration System (http://ameriflux.ornl.gov)
API	Application Program Interface (https://en.wikipedia.org/wiki/Application_programming_interface)
ARM	Atmospheric Radiation Measurement: The ARM Climate Research Facility is a DOE user facility that provides in situ and remote sensing observations to improve the understanding and climate model representations of clouds, aerosols, and their interactions with the Earth's surface (www.arm.gov).
ARMBE	ARM Best Estimate Data Products (http://www.arm.gov/instruments/armbe)
ASR	Atmospheric System Research (http://science.energy.gov/ber/research/cesd/atmospheric-system-research-program/)
CESD	Climate and Environmental Sciences Division (http://science.energy.gov/ber/research/cesd/)
CF	CF Conventions and Metadata (http://cfconventions.org)
CMIP5	Coupled Model Intercomparison Project, phase 5, sponsored by WCRP/WGCM, and related multi-model database planned for the IPCC AR5 (http://cmip-pcmdi.llnl.gov)
CMIP6	Coupled Model Intercomparison Project, phase 6, sponsored by WCRP/WGCM, and related multi-model database planned for the IPCC AR6 (http://cmip-pcmdi.llnl.gov)
Data Node	Internet location providing data access or processing (http://en.wikipedia.org/wiki/Node-to-node_data_transfer)
DOE	Department of Energy, the U.S. government entity chiefly responsible for implementing energy policy (http://www.doe.gov/)
EMSL	Environmental Molecular Science Laboratory (http://genomicscience.energy.gov/userfacilities/emsl.shtml)
EOS	NASA's Earth Observing System (http://eospso.nasa.gov)
ESGF	Earth System Grid Federation, led by LLNL, a worldwide federation of climate and computer scientists deploying a distributed multi-petabyte archive for climate science (http://esgf.llnl.gov)
ESM	Earth System Modelling (http://science.energy.gov/ber/research/cesd/earth-system-modeling-program/)
Esnet	Energy Sciences Network (https://www.es.net)
Globus	Provides high-performance, secure, and reliable data transfer, sharing, synchronization, and publication services for the science community (www.globus.org)
GridFTP	A high-performance, secure, reliable data transfer protocol optimized for high-bandwidth wide-area networks (http://toolkit.globus.org/toolkit/docs/latest-stable/gridftp/)
HPC	High Performance Computing (https://en.wikipedia.org/wiki/HPC)
IA	Integrated Assessment of Global Climate Change (http://science.energy.gov/ber/research/cesd/integrated-assessment-of-global-climate-change/)
ICNWG	International Climate Network Working Group, formed under the Earth System Grid Federation (ESGF), is to help set up and optimize network infrastructure for their climate data sites located around the world (http://icnwg.llnl.gov/)
INCITE	Innovative and Novel Computational Impact on Theory and Experiment program (http://www.doeleadershipcomputing.org/incite-program/)
LCF	DOE Leadership Compute Facilities (http://www.doeleadershipcomputing.org)

LLNL	Lawrence Livermore National Laboratory, sponsored by the DOE (https://www.llnl.gov/)
Metadata	Data properties, such as their origins, spatio-temporal extent, and format (http://en.wikipedia.org/wiki/Metadata)
NERSC	National Energy Research Scientific Computing Center, sponsored by the DOE (https://www.nersc.gov)
OLCF	Oak Ridge Leadership Computing Facility, sponsored by DOE (https://www.olcf.ornl.gov)
ORNL	Oak Ridge National Laboratory, sponsored by DOE (https://www.ornl.gov)
PNNL	Pacific Northwest National Laboratory, sponsored by DOE (http://www.pnnl.gov)
Py-ART	Python ARM Radar Toolkit, Python module containing a collection of weather radar algorithms and utilities (http://arm-doe.github.io/pyart/)
RGCM	Regional and Global Climate Modeling (http://science.energy.gov/ber/research/cesd/regional-and-global-modeling/)
SBR	Subsurface Biogeochemistry Research (http://science.energy.gov/ber/research/cesd/subsurface-biogeochemical-research/)
TES	Terrestrial Ecosystem Science (http://science.energy.gov/ber/research/cesd/terrestrial-ecosystem-science/)
UQ	Uncertainty quantification, method determining how likely a particular outcome is, given the inherent uncertainties or unknowns in a system (http://en.wikipedia.org/wiki/Uncertainty_quantification)
UV-CDAT	Ultrascale Visualization Climate Data Analysis Tools, provides access to large-scale data analysis and visualization tools for the climate modeling and observational communities (http://uvcdat.llnl.gov)
Web portal	A point of access to information on the World Wide Web (http://en.wikipedia.org/wiki/Web_portal)